# New computational methods and plant models for evolutionary genomics

Kevin D. Murray


A thesis submitted for the degree of
Doctor of Philosophy of
The Australian National University

May 6, 2019

This thesis was typeset in URW Garamond using the LaTeX typesetting system originally developed by Leslie Lamport, based on TeX created by Donald Knuth, and using pandoc, devised and maintained by John MacFarlane.

I declare that the research presented in this Thesis represents original work that I carried out during my candidature at the Australian National University, except for collaborator's contributions to multi-author papers incorporated in the Thesis, which are detailed in each chapter's prefix.

Kevin D Murray
kevin@kdmurray.id.au
May 6, 2019

From forest caves, and azure skies
We crashed upon this earth
The years, they passed
And so did we
But resistance would be bought

---

from *So Did We* – Isis

# Acknowledgements

This thesis would never have been possible without the support of an enormous number of people, too many to name individually. Nonetheless, you all have my gratitude.

Firstly, I must thank all those who have advised and mentored me over the last four years. Justin, thank you for everything, I would be neither as competent nor confident as a scientist without the advice, encouragement, and opportunities you've given me during and before my PhD. To Norman, who co-supervised my work on what became kWIP, thanks for the belief you had in my daft ideas, and for your enthusiasm and continuing friendship. To Rose, who co-supervised my recent work on *Eucalyptus*, thanks for your endless and cheerful help, and for your hospitality during my frequent visits to Armidale. To Sylvain, my late co-supervisor, your friendly advice influenced much of the work I present here, and you have been greatly missed. To my panel, Barry, Gavin, and Eric, thank you for all the encouragement and advice you've given. And an extra thanks to Barry for the opportunities you gave me as an undergraduate, without which I doubt I would have commenced a PhD.

I must also thank my co-authors and collaborators on the work I present here: Chrisfried, Cheng Soon, Jared, Pip, Steve, Riyan, Niccy, Jasmine, Helen, and anyone I've forgotten. Thanks also to the great communities in which I've worked, particularly the members of the Borevitz, Pogson, and Andrew labs, and my EEG and Plant Science PhD student cohort. Thanks to Tim Brown and the whole APPF team, who've been rewarding colleagues with whom to plumb the murky depths of phenomics data. Thanks to the international labs who I have visited and/or collaborated with: Titus Brown and lab (particularly Camille Scott, Luíz Irber, and Michael Crusoe), for your help with `khmer` and hospitality in Davis; Detlef Weigel and lab for hospitality in Tübingen; and Loren Rieseberg and lab for hospitality in Vancouver. Thanks to the developers of several tools, whose helpful advice enabled their use: Gideon Bradburd, Timothy Bilton, Johannes Köster, Titus Brown's lab, Vince Buffalo, Reed Cartwright, Paul Staab, and Jonas Meisner.

On a personal note, thank you to all my dear friends, whose friendship has somehow kept me approximately sane through the trials of a PhD. Thank you to my family, for getting me to the point that I could start a PhD, and for getting me through it. Most of all, thank you to Luisa, my dearest companion. You are incredible, and have made the last 5 years the best of my life.

Kevin Murray
May 6, 2019

# Thesis Abstract

This thesis is in the service of a greater understanding of the genetic basis of adaptive traits. Chapter 1 introduces background literature relevant to this thesis. Chapters 2, 3, and 4 develop novel methods and software for the analysis of genetic sequencing data. Chapter 5 details a large collaborative project to establish genetic resources in the model cereal Brachypodium, and perform a genome-wide association study for several agriculturally-relevant traits under two climate change scenarios. Chapter 6 investigates the spatial genetic patterns in two species of woodland eucalypt, and determines the landscape process that could be driving these patterns. Finally, Chapter 7 summarises these works, and proposes some areas of further study.

In Chapters 2 and 3, I develop methods that enable the analysis of Genotyping-by-sequencing data. Axe, a short read sequence demultiplexer, demultiplexes samples from multiplexed GBS sequencing datasets. I show Axe has high accuracy, and outperforms previously published software. Axe also tolerates complex indexing schemes such as the variable-length combinatorial indexes used in GBS data. Trimit and libqcpp (Chapter 3) implement several low-level sequence read quality assessment and control methods as a C++ library, and as a command line tool. Both these works have been published in peer-reviewed journals, and are used by numerous groups internationally.

In Chapter 4, I develop kWIP, a *de novo* estimator of genetic distance. kWIP enables rapid estimation of genetic distances directly from sequence reads. We first show kWIP outperforms a competing method at low coverage using simulations that mimic a population resequencing experiment. We propose and demonstrate several use cases for kWIP, including population resequencing, initial assessment of sample identity, and estimating metagenomic similarity. kWIP was published in PLoS Computational Biology.

In Chapter 5, I present the results of a large, collaborative project that surveys the global genetic diversity of the model cereal *Brachypodium*. We amass a collection of over 2000 accessions from the Brachypodium species complex. Using GBS and whole genome sequencing we identify around 800 accessions of the diploid *Brachypodium distachyon*, within which we

find extensive population structure and clonal families. Through population restructuring we create a core collection of 74 accessions containing the majority of the genetic diversity in the "A genome" sub-population. Using this core collection, we assay several phenotypes of agricultural interest including early vigour, harvest index and energy use efficiency under two climates, and dissect the genetic basis of these traits using a genome-wide association study (GWAS). This work has been published in Genetics; I am co-first author with Pip Wilson and Jared Streich, having lead many genomic analyses.

In Chapter 6, I perform a study of landscape genomic variation in two woodland eucalypt species. Using whole genome sequencing of around 200 individuals from around 20 localities of both *E. albens* and *E. sideroxylon*, I find incredible genetic diversity and low genome-wide inter-species differentiation. I find no support for strong discrete population structure, but strong support for isolation by (geographic) distance (IBD). Using generalised dissimilarity modelling, I further examine the pattern of IBD, and establish additional isolation by environment (IBE). *E. albens* shows moderately strong IBD, explaining 26% of the deviance in the genetic distance using geographic distance, and an additional 6% of the deviance is explained by incorporating environmental predictors (IBE). *E. sideroxylon* shows much stronger IBD, with 78% of the deviance explained by geography, and stronger IBE (12% additional deviance explained). This work will soon be submitted for publication.

# Contents

# Chapter 1

# Thesis Introduction

This thesis concerns the genomics of adaptation to environment in plants. The works that comprise this thesis address a wide range of questions in several systems, and establish novel methods to do so. They range from the development of improved algorithms for vital early stages of modern sequencing analysis pipelines, to a study of the landscape drivers of spatial genetic patterns in *Eucalyptus*. They span multiple generations of improvement in genome sequencing technology, highlighting the incredible pace of technological improvement this field is experiencing. This thesis also demonstrates that the leading edge of genomics can only be reached through the development of novel statistical and computational tools, and their collaborative application to large datasets.

## 1.1 Adaptation genomics: the genomic biology of populations, landscapes, and quantitative traits

The phenotype of an individual is the expression of its genome in the environment it experiences. Adaptation genomics centers on the triad of genotype, phenotype, and environment, and interrogates the processes linking these concepts (Bragg et al., 2015; Radwan Jacek and Babik Wiesaw, 2012; Stapley et al., 2010). Adaptation genomics asks a series of questions related through this triad: What potentially adaptive genetic diversity exists? What phenotypic diversity is expressed? How is this diversity distributed and filtered across the landscape? Which loci underpin variation in traits? Is there evidence for selection at these loci?

The fields of population, landscape, and quantitative genomics form the backbone of this thesis. These fields study intra- and interspecific variation from a variety of angles, and produce findings that can assist management of ecosystems, both natural and agricultural.

For example, an understanding of the environmental drivers of genetic variation over the landscape can help restore and manage natural ecosystems (Broadhurst et al., 2008; Hoffmann et al., 2015). In an agricultural context, dissecting the genetic basis of phenotypic traits enables precise selection of advantageous genotypes to improve both yield and resilience (Ainsworth and Ort, 2010; Fernie et al., 2006). The study of genetic diversity is essential to all these questions.

### 1.1.1 Genetic diversity

A genetic polymorphism is defined here as a locus at which mutation has given rise to at least two distinct allelic states that occur in multiple individuals. Over time the frequencies of these states may change (i.e., evolution), through either random drift or selection. Multi locus genetic diversity is a quantification of these polymorphisms among individuals within and between populations. It is the substrate of evolution controlling heritable phenotypic variation that is subject to natural selection underlying adaptation.

In most organisms, a proportion of genetic variation is spatially autocorrelated – that is, genetic variation is not randomly distributed over the geographic range of a species (Sokal and Oden, 1978a, 1978b). Genetic spatial autocorrelation arises through a variety of processes, both neutral and adaptive (Diniz-Filho et al., 2009). Many of the processes governing spatial autocorrelation of allele frequencies are not the result of any adaptive separation, for example outbreeding over limited distances or expansion from an ancestral refugia (Sokal and Oden, 1978b). Strong selection on specific traits may lead to reduced gene flow between differing environments (isolation by environment), perhaps because of unfit migrants (Wang and Bradburd, 2014).

### 1.1.2 Genetic isolation and differentiation

Genetic differentiation is due to neutral processes including drift, particularly in small populations, non-random mating, and non-random migration. Isolated subpopulations with reduced gene flow relative to drift will fix mutations that differentiate them (Hahn, 2018). Discrete population structure is the result of low gene flow between two (or more) subpopulations, relative to gene flow within each subpopulation. There are innumerable potential causes of reduction in gene flow, some geographic (e.g., separation by large swathes of intolerable habitat), and others non-geographic (e.g., divergence in flowering time leading to temporal isolation).

In contrast to discrete population structure, Isolation by Distance (IBD; Wright, 1943) is the observation that proximate individuals have higher relatedness than distant individuals; IBD is a ubiquitous pattern (Meirmans, 2012). IBD may vary in strength over the landscape and across the genome. In particular, the relationship between genetic and geographic distance need not be linear. While there are many specific causes of IBD, fundamentally it is the result of the probability of gene flow between two individuals being some function of their geographic separation (i.e., non-random mating). Technically, the pattern of IBD is the integral through time of this non-random mating, and the extent of non-random mating may vary through time. Additionally, certain demographic histories can result in a pattern of IBD, for example postglacial expansion from a refugia (Holliday et al., 2010; Meirmans, 2012).

Isolation by Environment (IBE; Wang and Bradburd, 2014) extends the concept of Isolation by Distance to environmental causes of non-random mating over the landscape. IBE occurs when individuals in dissimilar environments exchange less genetic material than individuals in similar environments, controlling for reduced gene flow caused by geographic separation. IBE can have many causes, including selection, reduced fitness of inter-environment migrants or hybrids, and biased dispersal. Importantly, although local adaptation at particular loci can ultimately lead to genome wide IBE, evidence of IBE is not evidence of selection: a variety of neutral processes can generate similar patterns (e.g., postglacial recolonisation *a la* IBD; Wang and Bradburd, 2014; Holliday et al., 2010). Typically, environmental distance is calculated from interpolated predictions (e.g. of climate; Xu and Hutchinson, 2013; Fick and Hijmans, 2017). Estimation of IBE assumes the underlying modelled environmental variables are accurate throughout the life of the individual — a tenuous assumption for certain variables in the case of long-lived organisms. Discriminating between patterns that are purely neutral, neutral but are side effects of selection (i.e. linked selection), and patterns that are a direct result of selection is challenging, and typically requires orthogonal information (Holliday et al., 2010; Meirmans, 2012; Wang and Bradburd, 2014). For instance, evidence of reduced fitness of inter-environmental immigrants can confirm local adaptation (Wang and Bradburd, 2014; e.g. in Keller et al., 2011), but do not get at the underlying genetic basis. Such experiments (e.g. provenance trials) are expensive and time consuming, especially in trees. Therefore, I make no attempt to do so in this thesis, testing only for patterns of correlation between geographic/environmental distance and genetic distance.

### 1.1.3 Conservation: migration of ecotypes vs increasing adaptive potential

The global climate is changing, and entire ecosystems are being lost through a variety of human activities (e.g. deforestation). One application of finding genetic isolation by environment and geography is to assist conservation and revegetation efforts at a given locality (Broadhurst et al., 2008; Supple et al., 2018). However, looking forward, local adaptation to (past) environment does not necessarily indicate suitability to the future environment at a given locality (Wogan and Wang, 2018). Even when models of IBD and IBE are used to predict genomic suitability from predicted future climate data, humility regarding model projections should guide recommendations. This is especially true given predictions suggest that not only will environmental means shift with climate change, but variances will increase (Thornton et al., 2014). This suggests that a broad understanding of the landscape processes governing spatial patterns of genetic variation is more urgent than discovering locally-adapted loci in each habitat, especially given limited resources (Kardos and Shafer, 2018). In any case, a focus on preserving adaptive potential is key to any restoration and conservation works, rather than attempts to discover "perfectly adapted germplasm" for revegetation (Broadhurst et al., 2008; Hoffmann et al., 2015; Weeks et al., 2011).

## 1.2 Discovering genetic variation

Virtually all modern studies of genetics now use either short or long read high throughput sequencing (Mardis, 2008; Metzker, 2010). Such methods generate truly staggering quantities of data within which genetic variation must be discovered (Pfeifer, 2017). Many molecular and computational methods exist to discover genetic variation, of which I will discuss those pertinent to the work I present in this thesis. Additionally, I will discuss the fact that off-the-shelf molecular and computational tools are often not suited to non-human experiments, mandating the development of new tools as part of a particular biological project.

Given sequencing machines do not provide full diploid genome sequences directly, genetic variation is typically discovered as variation among samples relative to a reference genome (Li, 2011; Pfeifer et al., 2014). This process, termed variant calling, statistically integrates all data across samples to determine if each position in the reference genome is variable, and subsequently to determine the most likely genotype at each locus for each individual. A variety of computational and statistical approaches have been developed to perform this task, for example mpileup (Li, 2011), freebayes (Garrison and Marth, 2012) and GATK's Halplotype-

Caller (DePristo et al., 2011; McKenna et al., 2010). These algorithms were largely designed to work with relatively high coverage (e.g. 15 fold), and with a reference genome that is not very diverged from samples of interest. Where these assumptions are not met, various errors and/or biases may be introduced, both in genotype calls and subsequent inferences based upon them (Brandt et al., 2015; Han et al., 2014; Nielsen et al., 2011). To combat these issues, various methods that aim to preserve uncertainty inherent in less than saturating sequencing coverage. This uncertainty can be incorporated into typical population genetic inference, e.g. ANGSD (Fumagalli et al., 2013; Korneliussen et al., 2014, 2013). These methods estimate the most likely values of the statistic of interest (e.g. calculation of population genetic statistics, or inter-sample distances) given the sequence data, rather than given the called genotypes. This difference allows these methods to improve the accuracy and reduce the bias of their estimates, particularly on low-coverage data (Fumagalli, 2013; Korneliussen et al., 2014, 2013).

Necessarily, discovering genetic variation involves integrating across many samples. However, even with methods designed to operate on low coverage data, reliably assaying a sample's genotype requires a significant sequencing cost. Given limited budgets, this presents a trade-off: more samples, with less data per sample, or fewer samples with more data? In nearly all cases, more samples with a modest amount of data represents the best statistical power for some fixed investment (Fumagalli, 2013; Kliebenstein, 2012; Li et al., 2011; Pasaniuc et al., 2012).

### 1.2.1 Dimensions of Genomic Complexity

Genomic complexity was considered within a reference haploid genome, measured perhaps by genome size, repeat content or ploidy. With population re-sequencing, there is another axis to consider, that of genetic diversity across samples. In this dimension, genetic diversity could be considered to be the total or average pairwise diversity among samples in a population. The units of this diversity axis would be metrics such as $\pi$, population structure and $F_{ST}$, or the extent of linkage disequilibrium (LD; Hahn, 2018). In studies that aim to determine functional genetic variation, both dimensions of genome complexity need to be considered as they determine the experimental design and cost. As both samples and polymorphisms show statistical dependence (i.e. structure and LD), missing data can be imputed across related individuals and along the genome (Browning and Browning, 2007; Howie et al., 2009; Marchini and Howie, 2010). This can sometimes dramatically reduce costs but can also limit resolution and power (fig. 1.1).

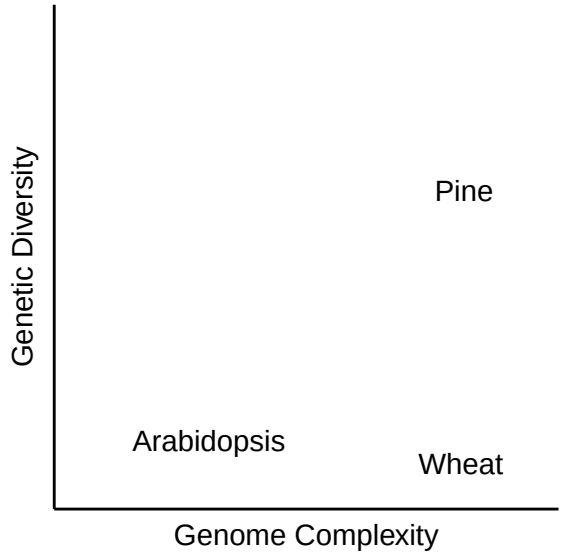



**Figure 1.1: Dimensions of genomic complexity.** In many genomic studies (particularly genome assembly), genome complexity describes the size and complexity of a single haploid genome. In studies of evolutionary genetics, we must also consider an orthogonal axis: genetic variability. The size and complexity of a single haploid genome determines the cost of reliably assaying the whole genome sequence of one individual, however imputation can leverage statistical dependence between loci to reduce the effective cost of sequencing a population. Thus, the total cost of genotyping a population in a resequencing study is some combination of both genomic complexity and genetic variation (specifically LD).

Regardless of the specific amount of structure and LD in the sampled genomes, a reference genome is either required or extremely useful (Pfeifer, 2017; although see e.g. Audano et al., 2018). In particular, polymorphism-wise analyses like GWAS require a reference genome to order markers, and to anchor them to genes. Ordered markers are also crucial for inference of some population metrics, particularly those whose currency is haplotypes (e.g. selective sweep detection; Hahn, 2018; Pavlidis et al., 2013; discovery of introgressed segments; Bragg et al., 2015). Thankfully, modern long-read sequencing methods and long read-specific assemblers make assembly of at least draft genomes achievable on limited budgets of time and money (e.g. Michael et al., 2018). I have also developed non-reference approaches to look at sample relatedness (Chapter 4; Murray et al., 2017) which can then prioritize selection of distantly related samples for reference genome assembly.

### 1.2.2 Specifics of sequencing methods

Until relatively recently, it remained too expensive for most researchers to sequence whole genomes with sufficient sample size for quantitative study of genetic variation. Reduced representation sequencing methods aimed to reduce sequencing costs by assaying only a fraction of the genome. A large range of such methods exists, employing various molecular methods to subset the genome (Baird et al., 2008; Blumenstiel et al., 2010; Elshire et al., 2011; Faircloth et al., 2015; Morris et al., 2011; Peterson et al., 2012; Smith et al., 2014). One such method employed in this thesis is Genotyping by Sequencing (GBS; Elshire et al., 2011). GBS uses restriction enzymes to fragment each sample's genome, and sequencing libraries are created only for genome regions near these restriction sites. Then, many hundreds of samples may be multiplexed and sequenced on one sequencing run, drastically reducing the sequencing cost per sample (Elshire et al., 2011). While such data is of acceptable quality and quantity for most genome-average analyses (e.g. genetic distance, detecting population structure), it often assays an insufficient fraction of the genome to be used in many genome-wide polymorphism-wise analyses (unless LD is very extensive). Additionally, GBS data is often plagued with a large proportion of missing data, a lack of reproducibility, strong batch effects, and strong allelic bias (Bilton et al., 2018; Lowry et al., 2017).

In contrast to GBS, whole-genome shotgun sequencing provides data on (nearly) all DNA present in each sample. In such methods, sequencing libraries are created from short fragments of DNA from approximately random genome positions (including organellar and other genomes present in a sample; Bentley et al., 2008). As a consequence, these methods require significantly more sequencing per sample than reduced representation data, but

generally produce datasets of higher quality, and assay (nearly) all polymorphisms present in some population (Fuentes-Pardo and Ruzzante, 2017; Lowry et al., 2017). With the advent of cheaper sequencing and development of cost-efficient library preparation (Jones et al., 2018), WGS became economic at increasing scales, to the point where WGS costs are now approximately equal the cost of GBS at the beginning of my PhD.

### 1.2.3 Computational method development

The development of robust and efficient computational methods often lags behind the development of ever more efficient molecular methods. Therefore, to use cutting edge molecular methods, one must often develop computational tools specific to some dataset or protocol. This is especially true in adaptation genomics, where the experimental designs (such as those discussed below) diverge significantly from those that many "off-the-shelf" analysis methods expect. The methods required range from smaller utilities to efficiently solve problems with known solutions, to large programs implementing entirely new metrics or algorithms.

While there are only a few high-level steps in any analysis (e.g. alignment to reference, variant calling), each high-level task is often performed by several smaller tools. In many cases, these tools are application- or datatype-specific. The use of custom, cost-conscious molecular protocols in the chapters of this thesis required the development of certain tools. One example is the requirement for improved demultiplexing software for GBS data. Increasing sequencing capacity lead to increasing amounts of sample pooling, in the case of GBS using combinatorial, paired end, indexing. However, at the time, no methods were able to de-multiplex this data accurately and efficiently. This required the development of tools to perform this crucial early stage of GBS data analysis (Chapters 2 and 3).

As outlined above, discovering genetic variation is computationally intensive, and typically requires a reference genome. However, at least in the early stages of analysis, one is simply interested in the genetic similarity of samples. Therefore, efficient methods to rapidly estimate of genetic similarity directly from sequencing data are required. Such methods have the potential to estimate genetic distances without reference genome bias, and to verify that individuals belong to the correct genetic lineage before conclusions are drawn using mislabelled, or misidentified samples. Importantly, such methods would be sequencing-method agnostic, unlike reference-free methods designed to work only with reduced representation methods.

Alignment-free sequence comparison algorithms satisfy many of these requirements. Alignment-free methods generally operate by decomposing sequences (or sequence data)

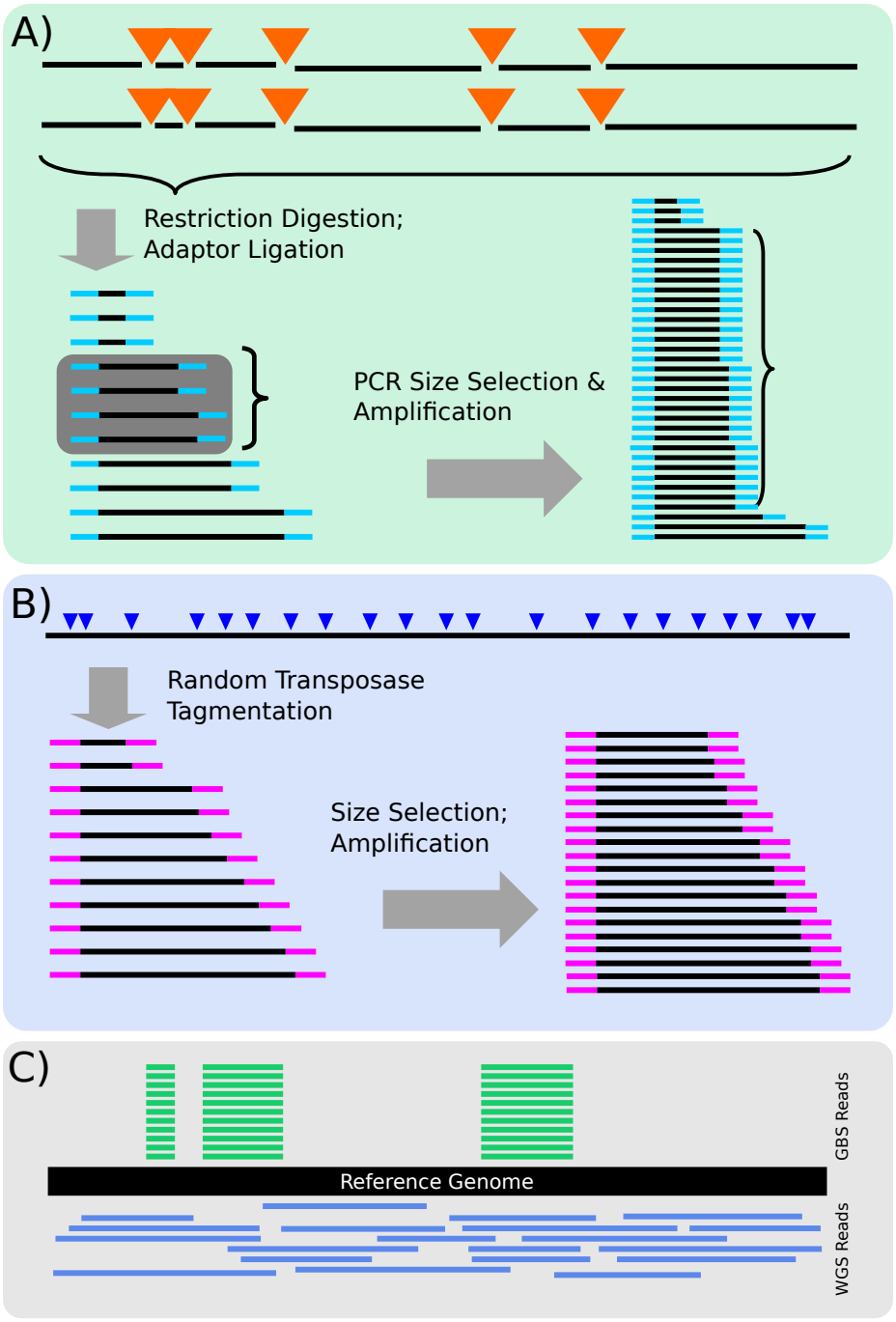



**Figure 1.2: Overview of GBS and WGS sequencing methods and data.** A) GBS sequencing protocol: briefly, genomic DNA is digested with restriction enzymes, adaptors are ligated, and PCR used to size-select and amplify libraries. B) WGS sequencing protocol: briefly, adaptors are directly transposed into genomic DNA at random positions, amplified with PCR, and size-selected using electrophoresis. C) GBS and WGS data aligned to a reference genome. GBS data aligns to quantised positions, and has higher coverage at covered positions. WGS data aligns approximately uniformly across the genome, and has lower coverage for a given sequencing volume.

into kmers, i.e. substrings of length k (e.g. Song et al., 2014; Forêt et al., 2009; Sims et al., 2009; Tang et al., 2014). Recently, several algorithms enabling *de novo* sequence comparison have been published, generally attempting to reconstruct phylogenetic relationships from sequencing reads. `Spaced` (Leimeister et al., 2014; Morgenstern et al., 2015) uses the Jensen-Shannon distance on spaced kmers for accurate phylogenetic reconstruction. `Cnidaria` (Aflitos et al., 2015) and `AAF` (Fan et al., 2015) use the Jaccard distance to estimate phylogenetic distances. `Mash` (Ondov et al., 2016) uses a MinHash approximation of Jaccard distance to the same effect but with extreme computational efficiency. My contribution, kWIP (Chapter 4) uses a weighted inner product to determine distance and is well suited for within species differentiation.

## 1.3 Experimental designs to uncover drivers of spatial genetic diversity

Landscape genomic studies typically form two phases, an exploratory descriptive phase followed by one or more targeted, hypothesis-driven experiments guided by patterns uncovered in the first phase (Bragg et al., 2015). In each phase, many samples are obtained, across the relevant geographic and environmental range (i.e. initially the whole range of the study species, then across specific environmental gradients relating to hypotheses under consideration; Bragg et al., 2015; Wang and Bradburd, 2014). One then employs population-wide sequencing to find genetic variation, either using reduced-representation or whole-genome methods.

Once genetic variation has been ascertained, the descriptive phase seeks to establish genome-wide background patterns of gene flow and isolation across the landscape. There are many approaches to this task. Several methods model genetic distance (or similarity, e.g. allelic covariance) as a function of geographic and/or environmental distances. Traditionally, methods like mantel and partial mantel tests, and multiple regression on matrices were employed (Legendre and Legendre, 2012; Lichstein, 2007; Mantel, 1967; Smouse et al., 1986), however in some scenarios these have undesirable statistical properties, including a high false positive rate (Guillot and Rousset, 2013). More modern methods address some of these concerns, including generalised dissimilarity modelling (GDM; Ferrier et al., 2002, 2007), and BEDASSLE (Bradburd et al., 2013). All these methods establish the extent to which genetic distances (and therefore underlying allele frequencies) are spatially and environmentally autocorrelated as a result of non-random gene flow. Importantly,

these models of genome-wide average distance are overwhelmingly influenced by neutral segregating mutations, and are not in themselves evidence of any adaptive process, though selection may influence gene flow, and generate this pattern (Wang and Bradburd, 2014).

The subsequent hypothesis driven analyses could take many forms. One may wish to find individual loci associated with specific environmental characters, amounting to correlation of allele frequency with some environmental gradient. Therefore, specific tests are required, for example BayEnv (Günther and Coop, 2013) or latent-factor mixed models (Stucki et al., 2017). Such polymorphism-wise analyses are best performed using whole-genome sequencing data, to ensure all genetic variation is tested.

Complementary experiments examine environmentally adaptive trait variation, likely filtered by the environment, to dissect the genetic basis of these traits using GWAS. This would involve growth of individuals from across some environmental gradient in common conditions, and phenotyping of traits relevant to environmental gradients of interest (e.g. water use efficiency along an aridity gradient). Differences in fitness could also be directly tested, perhaps using reciprocal transplants along some environmental gradient to test for local adaptation. In all these experiments, one must correct not only for any population structure (*a la* GWAS), but also for the genome-wide background isolation by distance and environment that would cause neutral loci to correlate with environmental gradients or trait variation. For this reason, the power of such analyses is inversely proportional to the strength of genome-wide IBE and IBD and population structure. Chapter 6 presents a descriptive analysis of the patterns of IBD and IBE, and does not perform any of these more targeted, hypothesis-driven experiments.

## 1.4 Experimental designs for uncovering genetic basis of quantitative traits

Traits are heritable characters of an individual, and quantitative traits are the subset of traits whose values form an approximately normal distribution when quantified in some population (Lynch and Walsh, 1998). Uncovering the genetic basis of such traits involves statistical association of trait values with genetic variation across the genome. The strength of this statistical association is assessed at each polymorphic locus. Loci where genetic variation and trait variation appear correlated are considered a quantitative trait locus (QTL). This association is purely statistical without further investigation of QTL, perhaps in the form of functional study of genes contained within the QTL of interest. In particular, in most studies

the polymorphisms underlying QTL are predominantly not causal, rather their allelic state is correlated with that of causal mutations (linkage disequilibrium; LD), and causal mutations themselves may not even be assayed. These studies can be conducted in natural populations (i.e. genome-wide association studies; GWAS) or within artificially-created mapping populations, e.g. recombinant inbred lines or nested association mapping populations (often termed QTL mapping).

In species with low outbreeding rates, finding sufficient genetic variation is the hard part of GWAS. To do so, one typically sequences many thousand individuals, and reduces this to a core set that maximises genetic diversity while minimising population structure and other confounding factors; an approach termed "population restructuring" (Brachi et al., 2011). At least for this initial restructuring, whole-genome data is not required and studies typically use a cost-effective reduced representation sequencing approach (Brachi et al., 2011; Elshire et al., 2011). However, merely choosing a subset of samples is not always sufficient to create a powerful population. One can obtain more diverse and less structured populations either through targeted re-sampling of diverse regions, or the creation of artificial mapping populations that release genetic diversity from background structure (Brachi et al., 2011). Once a diverse population has been obtained, it can then be sequenced using whole genome sequencing, providing polymorphism data required for a GWAS.

In contrast, finding diversity in outbred species is less challenging than accurately assaying it. Outbred species tend to have high genetic diversity, with very large numbers of polymorphisms, and often relatively less correlation between these polymorphisms (i.e. low LD; Nybom and Bartish, 2000). In such cases, population restructuring is less likely to be required, and whole-genome sequencing may be used immediately (Brachi et al., 2011). However, very large sample sizes are often required to achieve a statistically powerful polymorphism-wise association study, especially when the number of polymorphisms is very large (Pfeiffer and Gail, 2003; Purcell et al., 2003).

Regardless of the genetic diversity and the population genetics of the sample set at hand, plants must be grown and traits of interest must be phenotyped, typically using automated high-throughput phenotyping (Brown et al., 2014). Plants may be grown in laboratory conditions, with the downside that their growth conditions are unrealistic, with obvious consequences on the expression of traits. To combat this, one may grow plants in lab growth conditions that attempt to mimic the basic environmental characters of some region (e.g. light intensity and spectral quality, temperature, day length). Genotype-environment interactions may be investigated, with plants grown under multiple environments, and genetic basis of

some trait determined independently in each condition.

## 1.5 Biological case studies

### 1.5.1 Brachypodum

The genus *Brachypodium* is collectively an established model system of grass (and in particular cereal) biology, and is phylogenetically situated among the world's major cereal crops (Draper et al., 2001; reviewed in Kellogg, 2015). *B. distachyon* has received particular attention, with its small, diploid genome, rapid life cycle, and small stature making it a tractable organism for laboratory study. The development of a model cereal is of particular interest given the often very large, complex genomes, long life cycles, large stature of crop species, and the large divergence, in phenotypic and phylogenetic terms, between these crops and the established model dicot *Arabidopsis thaliana* (Brkljacic et al., 2011). Significant variation in ploidy and genome architecture exists in *Brachypodium*, for example the common *B. hybridum* is an allo-tetraploid (2n=4x) between *B. distachyon* and *B. staceii* (Catalán et al., 2012). Such ploidy variation enhances the utility of this species complex as a model of cereal evolution, as many cereals have similar variation in ploidy and genome architecture. Studies of genetic diversity in *B. distachyon* found moderate genetic variation, confounded by strong structure and inbreeding, and extensive migration with little recombination (Filiz et al., 2009; Vogel et al., 2009; reviewed in Scholthof et al., 2018).

Efforts to establish *Brachypodium* as a model for cereal quantitative genetics are ongoing, with a *B. distachyon* reference genome assembly first released in 2010 (Bd21; The International Brachypodium Initiative, 2010), and an independent assembly of Bd21-3 available prepublication from JGI Phytozome release 11. The first sizeable collection of wild accessions and associated phenotypes was released in 2014 (Tyler et al., 2014). Short satellite repeat (SSR) markers and a collection of inbred lines were developed from samples of Turkish origin (a diversity hotspot; Vogel et al., 2009). Collections of hundreds of accessions were developed, primarily as part of the USDA's Germplasm Resources Information Network (Scholthof et al., 2018). However, at the outset of the collaborative project I present in Chapter 5, no global collections of a suitable size for GWAS were publicly available.

### 1.5.2 Eucalyptus

The genus *Eucalyptus* comprises more than 800 described taxa, with natural distributions restricted to Australia and surrounding tropical islands. Eucalypts are the dominant and keystone tree species in most Australian habitats, and some species are hardwood forestry species of global significance. Eucalypts range in form from large shrubs to the world's tallest angiosperm ("Centurion", a *E. regnans* in Tasmania that recently reached 100 metres; National Register of Big Trees, 2013). Thornhill et al. (2015) estimate the age of the genus *Eucalyptus* to be approximately 70 My. The genus is classified into three main subgenera which each contain a hierarchical grouping of species into series and series to sections (Pryor and Johnson, 1971). For example, Chapter 6 concerns two species, *E. albens* and *E. sideroxylon*, from two different series (*buxaeles* and *melliodoriae*) within the section Adnataria of the subgenus Symphiomyrtus. Many species within *Eucalyptus* readily hybridise, often forming fertile offspring and hybrid swarms (e.g. Pryor, 1953). Eucalypts are generally pollinated by generalist insect or vertebrate pollinators, and preferentially outcross (Potts and Gore, 1995). Broadly, previous genomic studies of widespread eucalypt species reveal very high genetic diversity, and low population structure and isolation by distance and/or environment (e.g. Potts and Jordan, 1994; Bloomfield et al., 2011; Gauli et al., 2014; Griffin et al., 1987; Supple et al., 2018).

## 1.6 Chapter summaries

This thesis presents work that views evolutionary genetics from a wide variety of angles. Chapters 2 and 3 describe the development of new computational methods to analyse GBS data. Chapter 4 describes kWIP, an estimator of genetic distance that operates directly on short read sequencing data. Chapter 5 describes a large effort to establish genetic resources for GWAS in the model cereal Brachypodium, and uses these resources for a GWAS on agronomic traits under simulated climate change. Chapter 6 presents recent work that seeks to identify the landscape drivers of genetic divergence and diversity in two woodland eucalypt species. Chapter 7 summarises these works, and proposes several avenues for further investigation. Chapters 2, 3, 4, and 5 outline works that have been accepted in peer reviewed journals, and accepted articles are replicated in this thesis for convenience.

## 1.7 References


Aflitos SA, Severing E, Sanchez-Perez G, Peters S, de Jong H, de Ridder D. **2015**. "Cnidaria: Fast, reference-free clustering of raw and assembled genome and transcriptome NGS data." *BMC Bioinformatics* **16**:352. doi:10.1186/s12859-015-0806-7

Ainsworth EA, Ort DR. **2010**. "How Do We Improve Crop Production in a Warming World?" *Plant Physiology* **154**:526. doi:10.1104/pp.110.161349

Audano PA, Ravishankar S, Vannberg FO, Berger B. **2018**. "Mapping-free variant calling using haplotype reconstruction from k-mer frequencies." *Bioinformatics* **34**:1659–1665. doi:10.1093/bioinformatics/btx753

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. **2008**. "Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers." *PLoS ONE* **3**:e3376. doi:10.1371/journal.pone.0003376

Bentley DR et al. **2008**. "Accurate whole human genome sequencing using reversible terminator chemistry." *Nature* **456**:53–59. doi:10.1038/nature07517

Bilton TP, Schofield MR, Black MA, Chagné D, Wilcox PL, Dodds KG. **2018**. "Accounting for Errors in Low Coverage High-Throughput Sequencing Data When Constructing Genetic Maps Using Biparental Outcrossed Populations." *Genetics* **209**:65–76. doi:10.1534/genetics.117.300627

Bloomfield JA, Nevill P, Potts BM, Vaillancourt RE, Steane DA. **2011**. "Molecular genetic variation in a widespread forest tree species Eucalyptus obliqua (Myrtaceae) on the island of Tasmania." *Aust J Bot* **59**:226–237. doi:10.1071/BT10315

Blumenstiel B et al. **2010**. "Targeted Exon Sequencing by In-Solution Hybrid Selection." *Current Protocols in Human Genetics* **66**:18.4.1–18.4.24. doi:10.1002/0471142905.hg1804s66

Brachi B, Morris GP, Borevitz JO. **2011**. "Genome-wide association studies in plants: The missing heritability is in the field." *Genome Biol* **12**:232. doi:10.1186/gb-2011-12-10-232

Bradburd GS, Ralph PL, Coop GM. **2013**. "Disentangling the effects of geographic and ecological isolation on genetic differentiation." *Evolution* **67**. doi:10.1111/evo.12193

Bragg JG, Supple MA, Andrew RL, Borevitz JO. **2015**. "Genomic variation across landscapes: Insights and applications." *New Phytol* **207**:953–967. doi:10.1111/nph.13410

Brandt DYC, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. **2015**. "Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Project Phase I Data." *G3* **5**:931–941. doi:10.1534/g3.114.015784

Brkljacic J et al. **2011**. "Brachypodium as a Model for the Grasses: Today and the Future." *Plant Physiology* **157**:3–13. doi:10.1104/pp.111.179531

Broadhurst LM, Lowe A, Coates DJ, Cunningham SA, McDonald M, Vesk PA, Yates C. **2008**. "Seed supply for broadscale restoration: Maximizing evolutionary potential." *Evolutionary Applications* **1**:587–597. doi:10.1111/j.1752-4571.2008.00045.x

Brown TB et al. **2014**. "TraitCapture: Genomic and environment modelling of plant phenomic data." *Current Opinion in Plant Biology* **18**:73–79. doi:10.1016/j.pbi.2014.02.002

Browning SR, Browning BL. **2007**. "Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering." *The American Journal of Human Genetics* **81**:1084–1097. doi:10.1086/521987

Catalán P et al. **2012**. "Evolution and taxonomic split of the model grass Brachypodium distachyon." *Ann Bot* **109**:385–405. doi:10.1093/aob/mcr294

DePristo MA et al. **2011**. "A framework for variation discovery and genotyping using next-generation DNA sequencing data." *Nat Genet* **43**:491–498. doi:10.1038/ng.806

Diniz-Filho JAF, Nabout JC, de Campos Telles MP, Soares TN, Rangel TFLVB. **2009**. "A review of techniques for spatial modeling in geographical, conservation and landscape genetics." *Genet Mol Biol* **32**:203–211. doi:10.1590/S1415-47572009000200001

Draper J, Mur LAJ, Jenkins G, Ghosh-Biswas GC, Bablak P, Hasterok R, Routledge APM. **2001**. "Brachypodium distachyon. A New Model System for Functional Genomics in Grasses." *Plant Physiology* **127**:1539–1555. doi:10.1104/pp.010196

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. **2011**. "A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species." *PLoS ONE* **6**:e19379. doi:10.1371/journal.pone.0019379

Faircloth BC, Branstetter MG, White ND, Brady SG. **2015**. "Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera." *Molecular Ecology Resources* **15**:489–501. doi:10.1111/1755-0998.12328

Fan H, Ives AR, Surget-Groba Y, Cannon CH. **2015**. "An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data." *BMC Genomics* **16**:522. doi:10.1186/s12864-015-1647-5

Fernie AR, Tadmor Y, Zamir D. **2006**. "Natural genetic variation for improving crop quality." *Current Opinion in Plant Biology*, Genome studies and molecular genetics: Part 1: Model legumes / edited by Nevin D Young and Randy C Shoemaker; Part 2: Maize genomics / edited by Susan R Wessler. Plant biotechnology / edited by John Salmeron and Luis R Herrera-Estrella **9**:196–202. doi:10.1016/j.pbi.2006.01.010

Ferrier S, Drielsma M, Manion G, Watson G. **2002**. "Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level

modelling." *Biodiversity and Conservation* **11**:2309–2338. doi:10.1023/A:1021374009951

Ferrier S, Manion G, Elith J, Richardson K. **2007**. "Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment." *Diversity and Distributions* **13**:252–264. doi:10.1111/j.1472-4642.2007.00341.x

Fick SE, Hijmans RJ. **2017**. "WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas." *International Journal of Climatology* **37**:4302–4315. doi:10.1002/joc.5086

Filiz E, Ozdemir BS, Budak F, Vogel JP, Tuna M, Budak H. **2009**. "Molecular, morphological, and cytological analysis of diverse Brachypodium distachyon inbred lines." *Genome* **52**:876–890. doi:10.1139/G09-062

Forêt S, Wilson SR, Burden CJ. **2009**. "Characterizing the D2 Statistic: Word Matches in Biological Sequences." *sagmb* **8**:1–21. doi:10.2202/1544-6115.1447

Fuentes-Pardo AP, Ruzzante DE. **2017**. "Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations." *Mol Ecol* n/a–n/a. doi:10.1111/mec.14264

Fumagalli M. **2013**. "Assessing the Effect of Sequencing Depth and Sample Size in Population Genetics Inferences." *PLOS ONE* **8**:e79667. doi:10.1371/journal.pone.0079667

Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A, Nielsen R. **2013**. "Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data." *Genetics* **195**:979–992. doi:10.1534/genetics.113.154740

Garrison E, Marth G. **2012**. "Haplotype-based variant detection from short-read sequencing."

Gauli A, Steane DA, Vaillancourt RE, Potts BM. **2014**. "Molecular genetic diversity and population structure in Eucalyptus pauciflora subsp. Pauciflora (Myrtaceae) on the island of Tasmania." *Aust J Bot* **62**:175–188. doi:10.1071/BT14036

Griffin AR, Moran GF, Fripp YJ. **1987**. "Preferential Outcrossing in Eucalyptus regnans F. Muell." *Aust J Bot* **35**:465–475. doi:10.1071/bt9870465

Guillot G, Rousset F. **2013**. "Dismantling the Mantel tests." *Methods in Ecology and Evolution* **4**:336–344. doi:10.1111/2041-210x.12018

Günther T, Coop G. **2013**. "Robust Identification of Local Adaptation from Allele Frequencies." *Genetics* **195**:205–220. doi:10.1534/genetics.113.152462

Hahn MW. **2018**. "Molecular population genetics." New York : Sunderland, MA: Oxford University Press ; Sinauer Associates.

Han E, Sinsheimer JS, Novembre J. **2014**. "Characterizing Bias in Population

Genetic Inferences from Low-Coverage Sequencing Data." *Mol Biol Evol* **31**:723–735. doi:10.1093/molbev/mst229

Hoffmann A et al. **2015**. "A framework for incorporating evolutionary genomics into biodiversity conservation and management." *Climate Change Responses* **2**:1. doi:10.1186/s40665-014-0009-x

Holliday JA, Yuen M, Ritland K, Aitken SN. **2010**. "Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests." *Molecular Ecology* **19**:3857–3864. doi:10.1111/j.1365-294X.2010.04767.x

Howie BN, Donnelly P, Marchini J. **2009**. "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies." *PLOS Genetics* **5**:e1000529. doi:10.1371/journal.pgen.1000529

Jones A, Borevitz J, Warthmann N. **2018**. "Cost-conscious generation of multiplexed short-read DNA libraries for whole genome sequencing v1 (protocols.io.unbevan)." doi:10.17504/protocols.io.unbevan

Kardos M, Shafer ABA. **2018**. "The Peril of Gene-Targeted Conservation." *Trends in Ecology & Evolution* **33**:827–839. doi:10.1016/j.tree.2018.08.011

Keller SR, Soolanayakanahally RY, Guy RD, Silim SN, Olson MS, Tiffin P. **2011**. "Climate-driven local adaptation of ecophysiology and phenology in balsam poplar, Populus balsamifera L. (Salicaceae)." *Am J Bot* **98**:99–108. doi:10.3732/ajb.1000317

Kellogg EA. **2015**. "Brachypodium distachyon as a Genetic Model System." *Annual Review of Genetics* **49**:1–20. doi:10.1146/annurev-genet-112414-055135

Kliebenstein DJ. **2012**. "Exploring the Shallow End; Estimating Information Content in Transcriptomics Studies." *Front Plant Sci* **3**. doi:10.3389/fpls.2012.00213

Korneliussen TS, Albrechtsen A, Nielsen R. **2014**. "ANGSD: Analysis of Next Generation Sequencing Data." *BMC Bioinformatics* **15**:356. doi:10.1186/s12859-014-0356-4

Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R. **2013**. "Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data." *BMC Bioinformatics* **14**:289. doi:10.1186/1471-2105-14-289

Legendre P, Legendre L. **2012**. "Numerical ecology, Third English edition. ed," Developments in environmental modelling. Amsterdam: Elsevier.

Leimeister C-A, Boden M, Horwege S, Lindner S, Morgenstern B. **2014**. "Fast alignment-free sequence comparison using spaced-word frequencies." *Bioinformatics* btu177. doi:10.1093/bioinformatics/btu177

Li H. **2011**. "A statistical framework for SNP calling, mutation discovery, association

mapping and population genetical parameter estimation from sequencing data." *Bioinformatics* **27**:2987–2993. doi:10.1093/bioinformatics/btr509

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. **2011**. "Low-coverage sequencing: Implications for design of complex trait association studies." *Genome Res* **21**:940–951. doi:10.1101/gr.117259.110

Lichstein JW. **2007**. "Multiple regression on distance matrices: A multivariate spatial analysis tool." *Plant Ecol* **188**:117–131. doi:10.1007/s11258-006-9126-3

Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, Storfer A. **2017**. "Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation." *Mol Ecol Resour* **17**:142–152. doi:10.1111/1755-0998.12635

Lynch M, Walsh B. **1998**. "Genetics and analysis of quantitative traits." Sunderland, Mass: Sinauer.

Mantel N. **1967**. "The detection of disease clustering and a generalized regression approach." *Cancer Res* **27**:209–220.

Marchini J, Howie B. **2010**. "Genotype imputation for genome-wide association studies." *Nature Reviews Genetics* **11**:499–511. doi:10.1038/nrg2796

Mardis ER. **2008**. "The impact of next-generation sequencing technology on genetics." *Trends in Genetics* **24**:133–141. doi:10.1016/j.tig.2007.12.007

McKenna A et al. **2010**. "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data." *Genome Res* **20**:1297–1303. doi:10.1101/gr.107524.110

Meirmans PG. **2012**. "The trouble with isolation by distance." *Molecular Ecology* **21**:2839–2846. doi:10.1111/j.1365-294X.2012.05578.x

Metzker ML. **2010**. "Sequencing technologies — the next generation." *Nat Rev Genet* **11**:31–46. doi:10.1038/nrg2626

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. **2018**. "High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell." *Nature Communications* **9**:541. doi:10.1038/s41467-018-03016-2

Morgenstern B, Zhu B, Horwege S, Leimeister CA. **2015**. "Estimating evolutionary distances between genomic sequences from spaced-word matches." *Algorithms for Molecular Biology* **10**:5. doi:10.1186/s13015-015-0032-x

Morris GP, Grabowski PP, Borevitz JO. **2011**. "Genomic diversity in switchgrass (Panicum virgatum): From the continental scale to a dune landscape." *Molecular Ecology* **20**:4938–4952. doi:10.1111/j.1365-294X.2011.05335.x

Murray KD, Webers C, Ong CS, Borevitz J, Warthmann N. **2017**. "kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity." *PLOS Computational Biology* **13**:e1005727. doi:10.1371/journal.pcbi.1005727

National Register of Big Trees. **2013**. "Tree Register: Centurion." `https://www.nationalregisterofbigtrees.com.au/listing_view.php?listing_id=205`

Nielsen R, Paul JS, Albrechtsen A, Song YS. **2011**. "Genotype and SNP calling from next-generation sequencing data." *Nat Rev Genet* **12**:443–451. doi:10.1038/nrg2986

Nybom H, Bartish IV. **2000**. "Effects of life history traits and sampling strategies on genetic diversity estimates obtained with RAPD markers in plants." *Perspectives in Plant Ecology, Evolution and Systematics* **3**:93–114. doi:10.1078/1433-8319-00006

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. **2016**. "Mash: Fast genome and metagenome distance estimation using MinHash." *Genome Biology* **17**:132. doi:10.1186/s13059-016-0997-x

Pasaniuc B et al. **2012**. "Extremely low-coverage sequencing and imputation increases power for genome-wide association studies." *Nat Genet* **44**:631–635. doi:10.1038/ng.2283

Pavlidis P, ivkovi D, Stamatakis A, Alachiotis N. **2013**. "SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes." *Mol Biol Evol*. doi:10.1093/molbev/mst112

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. **2012**. "Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species." *PLoS ONE* **7**:e37135. doi:10.1371/journal.pone.0037135

Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. **2014**. "PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R." *Mol Biol Evol* msu136. doi:10.1093/molbev/msu136

Pfeifer SP. **2017**. "From next-generation resequencing reads to a high-quality variant data set." *Heredity* **118**:111–124. doi:10.1038/hdy.2016.102

Pfeiffer RM, Gail MH. **2003**. "Sample size calculations for population- and family-based case-control association studies on marker genotypes." *Genet Epidemiol* **25**:136–148. doi:10.1002/gepi.10245

Potts BM, Gore PL. **1995**. "Reproductive biology and controlled pollination of Eucalyptus-a review." University of Tasmania.

Potts BM, Jordan GJ. **1994**. "The Spatial Pattern and Scale of Variation in Eucalyptus globulus ssp Globulus: Variation in Seedling Abnormalities and Early Growth." *Aust J Bot* **42**:471–492. doi:10.1071/bt9940471

Pryor LD. **1953**. "Anther shape in Eucalyptus genetics and systematics." *Proceedings of the Linnean Society of New South Wales* **78**:43–48.

Pryor LD, Johnson LAS. **1971**. "A classification of the Eucalypts." Canberra: Australian National University.

Purcell S, Cherny SS, Sham PC. **2003**. "Genetic Power Calculator: Design of linkage and association genetic mapping studies of complex traits." *Bioinformatics* **19**:149–150.

Radwan Jacek, Babik Wiesaw. **2012**. "The genomics of adaptation." *Proceedings of the Royal Society B: Biological Sciences* **279**:5024–5028. doi:10.1098/rspb.2012.2322

Scholthof K-BG, Irigoyen S, Catalan P, Mandadi K. **2018**. "Brachypodium: A monocot grass model system for plant biology." *The Plant Cell* tpc.00083.2018. doi:10.1105/tpc.18.00083

Sims GE, Jun S-R, Wu GA, Kim S-H. **2009**. "Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions." *Proc Natl Acad Sci U S A* **106**:2677–2682. doi:10.1073/pnas.0813249106

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. **2014**. "Target Capture and Massively Parallel Sequencing of Ultraconserved Elements for Comparative Studies at Shallow Evolutionary Time Scales." *Syst Biol* **63**:83–95. doi:10.1093/sysbio/syt061

Smouse PE, Long JC, Sokal RR. **1986**. "Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence." *Syst Biol* **35**:627–632. doi:10.2307/2413122

Sokal RR, Oden NL. **1978a**. "Spatial autocorrelation in biology: 1. Methodology." *Biological Journal of the Linnean Society* **10**:199–228. doi:10.1111/j.1095-8312.1978.tb00013.x

Sokal RR, Oden NL. **1978b**. "Spatial autocorrelation in biology: 2. Some biological implications and four applications of evolutionary and ecological interest." *Biological Journal of the Linnean Society* **10**:229–249. doi:10.1111/j.1095-8312.1978.tb00014.x

Song K, Ren J, Reinert G, Deng M, Waterman MS, Sun F. **2014**. "New developments of alignment-free sequence comparison: Measures, statistics and next-generation sequencing." *Brief Bioinform* **15**:343–353. doi:10.1093/bib/bbt067

Stapley J et al. **2010**. "Adaptation genomics: The next generation." *Trends in Ecology & Evolution* **25**:705–712. doi:10.1016/j.tree.2010.09.002

Stucki S et al. **2017**. "High performance computation of landscape genomic models including local indicators of spatial association." *Mol Ecol Resour* **17**:1072–1089. doi:10.1111/1755-0998.12629

Supple MA, Bragg JG, Broadhurst LM, Nicotra AB, Byrne M, Andrew RL, Widdup A, Aitken NC, Borevitz JO. **2018**. "Landscape genomic prediction for restoration of a Eucalyp-

tus foundation species under climate change." *eLife* **7**:e31835. doi:10.7554/eLife.31835

Tang J, Hua K, Chen M, Zhang R, Xie X. **2014**. "A novel k-word relative measure for sequence comparison." *Computational Biology and Chemistry* **53, Part B**:331–338. doi:10.1016/j.compbiolchem.2014.10.007

The International Brachypodium Initiative. **2010**. "Genome sequencing and analysis of the model grass Brachypodium distachyon." *Nature* **463**:763–768. doi:10.1038/nature08747

Thornhill AH, Ho SYW, Külheim C, Crisp MD. **2015**. "Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny." *Molecular Phylogenetics and Evolution* **93**:29–43. doi:10.1016/j.ympev.2015.07.007

Thornton PK, Ericksen PJ, Herrero M, Challinor AJ. **2014**. "Climate variability and vulnerability to climate change: A review." *Glob Chang Biol* **20**:3313–3328. doi:10.1111/gcb.12581

Tyler L, Fangel JU, Fagerström AD, Steinwand MA, Raab TK, Willats WG, Vogel JP. **2014**. "Selection and phenotypic characterization of a core collection of Brachypodium distachyon inbred lines." *BMC Plant Biology* **14**:25. doi:10.1186/1471-2229-14-25

Vogel JP, Tuna M, Budak H, Huo N, Gu YQ, Steinwand MA. **2009**. "Development of SSR markers and analysis of diversity in Turkish populations of Brachypodium distachyon." *BMC Plant Biology* **9**:88. doi:10.1186/1471-2229-9-88

Wang IJ, Bradburd GS. **2014**. "Isolation by environment." *Mol Ecol* **23**:5649–5662. doi:10.1111/mec.12938

Weeks AR et al. **2011**. "Assessing the benefits and risks of translocations in changing environments: A genetic perspective." *Evol Appl* **4**:709–725. doi:10.1111/j.1752-4571.2011.00192.x

Wogan GOU, Wang IJ. **2018**. "The value of space-for-time substitution for studying fine-scale microevolutionary processes." *Ecography* **41**:1456–1468. doi:10.1111/ecog.03235

Wright S. **1943**. "Isolation by Distance." *Genetics* **28**:114–138.

Xu T, Hutchinson MF. **2013**. "New developments and applications in the ANUCLIM spatial climatic and bioclimatic modelling package." *Environmental Modelling & Software* **40**:267–279. doi:10.1016/j.envsoft.2012.10.003

# Chapter 2

# Axe: rapid, competitive sequence read demultiplexing using a trie

In this chapter, I describe a new method for demultiplexing short read sequencing data (e.g. Illumina). Axe efficiently matches the sequence of each read to a precomputed mapping of the expected index sequences (allowing for sequencing error), and outputs each sample's sequence data as an independent file. Axe was created to operate on Genotyping-by-Sequencing data, however it is compatible with any sequencing approach with in-read index sequences. I devised this algorithm, implemented it in the Axe program, and designed and executed the experiments that show Axe is at least as accurate as and far faster than several competing methods.

This chapter is published as an application note in Bioinformatics (2018; doi: 10.1093/bioinformatics/bty432). The senior author authorises the inclusion of this manuscript in my thesis.

*Bioinformatics*, 34(22), 2018, 3924–3925
doi: 10.1093/bioinformatics/bty432
Advance Access Publication Date: 1 June 2018
Applications Note



Sequence analysis

# Axe: rapid, competitive sequence read demultiplexing using a trie

**Kevin D. Murray* and Justin O. Borevitz**

ARC Centre of Excellence in Plant Energy Biology, Department of Plant Science, Research School of Biology, ANU, Canberra, Australia

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on March 5, 2018; revised on April 30, 2018; editorial decision on May 21, 2018; accepted on May 30, 2018

## Abstract

**Summary:** We describe a rapid algorithm for demultiplexing DNA sequence reads with in-read indices. Axe selects the optimal index present in a sequence read, even in the presence of sequencing errors. The algorithm is able to handle combinatorial indexing, indices of differing length and several mismatches per index sequence.

**Availability and implementation:** Axe is implemented in C, and is used as a command-line program on Unix-like systems. Axe is available online at https://github.com/kdmurray91/axe, and is available in Debian/Ubuntu distributions of GNU/Linux as the package axe-demultiplexer.

**Contact:** axe@kdmurray.id.au

**Supplementary information:** Supplementary data are available at *Bioinformatics* online

## 1 Introduction

The incredible yield of modern DNA sequencing technologies has enabled the multiplexing of DNA samples into a single sequencing unit. Multiplexing is achieved by the addition of short sequences (indices) to each molecule to be sequenced. When sequenced, these index sequences uniquely identify the sample to which a sequence read belongs. Many commercial protocols use platform specific features to add these DNA indices such that sequencing platforms can automatically demultiplex these samples. However, many custom sequencing protocols, including GBS (Elshire *et al.*, 2011), add indices which end users must themselves demultiplex. Combinatorial indexing schemes add independent index sequences to both pairs of a paired-end sequencing protocol, and samples are identified by the combination of these two index sequences [e.g. (Peterson *et al.*, 2012)].

Many sequencing read demultiplexers have been published. For example, both Flexbar (Dodt *et al.*, 2012) and the Fastx-toolkit's fastx_barcode_splitter.pl (Available at http://hannonlab.cshl.edu/fastx_toolkit/) accept single- and paired-end reads, however they cannot demultiplex combinatorial indices. AdapterRemoval (Schubert *et al.*, 2016) can demultiplex combinatorial indices, but cannot demultiplex indexes which differ in length. The same is true of DeML (Renaud *et al.*, 2015), which also uses a trie data structure. We developed axe to address these shortcomings.

## 2 Materials and methods

### 2.1 Algorithm

Axe matches the prefix of a sequence read against a pre-computed trie of index sequences. To do so, axe first calculates all sequences within a given Hamming distance (Hamming, 1950) of each index sequence. Axe then associates each of these sequences with its respective sample identifier using a double-array trie. Reads are demultiplexed by finding the read's longest prefix in the trie of (possibly mutated) index sequences, and assigning that read to its associated sample. This algorithm extends easily to combinatorial indexing, where two independent indices prefix each read of a read pair. The constant-time nature of these lookups allows Axe to remain rapid even with many thousand possible samples. Although this algorithm is agnostic as to which end of a sequencing read contains a index, only 5' (prefix) index demultiplexing is currently implemented.

### 2.2 Operation

To demultiplex sequence reads, one uses the command axe-demux. This command takes input reads as FASTQ or FASTA files which may contain single- or paired-end reads. Paired-end reads may be interleaved, and output reads can be written in any of these formats.

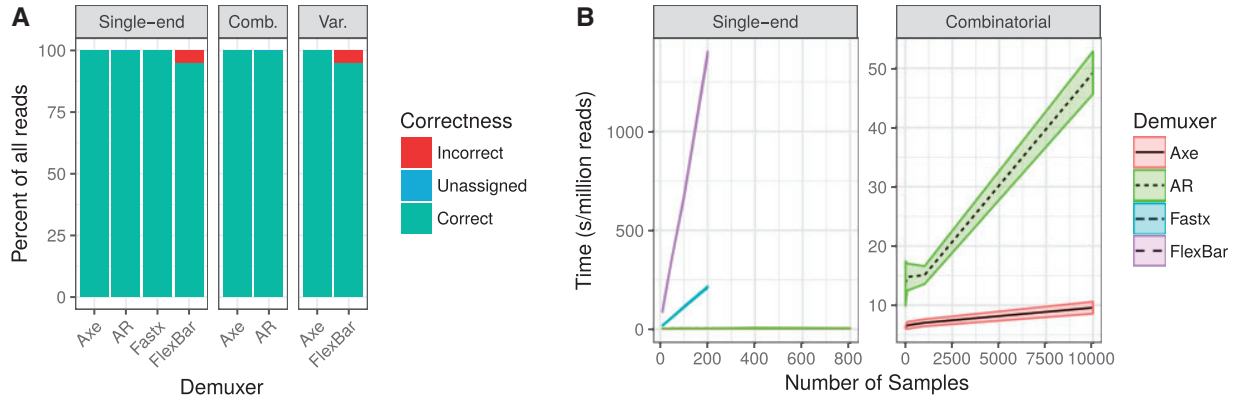



**Fig. 1.** (**A**) Accuracy of read assignment. Axe is able to perfectly demultiplex all reads, as is fastx. Only flexbar incorrectly assigns reads. Note: 'Comb.' refers to combinatorial index sets, and 'Var.' refers to index sets with variable length index sequences. (**B**) Computational performance of demultiplexers (seconds per million reads, mean ± SD). Axe is the fastest in all cases, closely followed by AdapterRemoval. fastx and flexbar are appreciably slower, especially when the number of indices is large

Axe is implemented in the C language, and is available at https://github.com/kdmurray91/axe. It may be built from source code on any modern POSIX operating system (including GNU/Linux and Mac OS X). The only dependencies not bundled with the source distribution are CMake and zlib. Axe is also available in the Debian and Ubuntu GNU/Linux distributions as the axe-demultiplexer software package.

## 3 Results

### 3.1 Demultiplexing accuracy and performance

We benchmark the speed and accuracy of axe, flexbar, AdapterRemoval and fastx_barcode_splitter.pl (hereafter 'fastx'). When demultiplexing read pairs with an index sequence on one read only (single-end), both axe and fastx are able to perfectly demultiplex all reads, with no error and with no reads left unassigned. AdapterRemoval fails to assign a minuscule proportion of reads, while flexbar mis-assigns several percent of reads (Fig. 1A). When demultiplexing combinatorially indexed read pairs, axe again demultiplexes all reads perfectly and AdapterRemoval fails to assign a small proportion. When demultiplexing reads with variable-length index sequences, axe performs perfectly, while flexbar mis-assigns several percent of reads. In all cases, axe is the fastest demultiplexer tested. AdapterRemoval performs several times slower than axe. fastx and flexbar perform hundreds of times slower than axe and AdapterRemoval (Fig. 1B). The methods underlying these experiments are available as online Supplementary Material.

### 3.2 Summary

Here, we implement a rapid and accurate algorithm for demultiplexing 5'-indexed reads. We show equal or improved accuracy and reduced computational cost compared to previous software developed to perform this task. In addition, more complex indexing schemes including combinatorial and/or variable length index sequences are supported. While in-read indexing is being phased out in some protocols, it persists in others such as GBS (Elshire *et al.*, 2011) and RNAseq using unique molecular identifiers (Kivioja *et al.*, 2012). Additionally, Axe's algorithm is applicable to demultiplexing out-of-read indexing schemes, though the implementation does not currently support this.

## Funding

This work was supported by the Australian Research Council Centre of Excellence in Plant Energy Biology [CE140100008], and was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government. KDM is supported by an Australian Government Research Training Program (RTP) Scholarship.

*Conflict of Interest*: none declared.

## References

Dodt,M. *et al.* (2012) FLEXBAR: flexible barcode and adapter processing for next-generation sequencing platforms. *Biology*, **1**, 895–905.

Elshire,R.J. *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Hamming,R.W. (1950) Error detecting and error correcting codes. *Bell Syst. Tech. J.*, **29**, 147–160.

Kivioja,T. *et al.* (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72–74.

Peterson,B.K. *et al.* (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, **7**, e37135.

Renaud,G. *et al.* (2015) deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics*, **31**, 770–772.

Schubert,M. *et al.* (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes*, **9**, 88.

# Chapter 3

# libqcpp: A C++14 sequence quality control library

This chapter outlines a software project that aimed to implement many next-gen sequencing quality control measures in a unified, reusable fashion. Libqcpp is a C++ library that presents an interface to several sequence quality control measures, including base quality measurement, read filtering, GBS-specific trimming, removal of low-coverage bases and quality reporting. Trimit is a command-line interface to this library. I designed and implement Trimit and Libqcpp, devised and tested several new quality control metrics, and wrote the software paper (including the online tutorials and use cases).

This paper was accepted for publication in the Journal of Open Source Software (2017; doi: 10.21105/joss.00232). The Journal of Open Source Software is a new, peer-reviewed journal for the dissemination of novel software and algorithms to a bioinformatics audience. JOSS papers are a summary of each software tool, and articles are accepted only when software meets several requirements: validation experiments are implemented in automated tests, use cases are presented in documentation and tutorials, and code is of a high quality. The senior author authorises the inclusion of this manuscript in my thesis.

The Journal of Open Source Software

# libqcpp: A C++14 sequence quality control library

**Kevin D Murray**[1] **and Justin O Borevitz**[1]

**1** ARC Centre of Excellence in Plant Energy Biology, The Australian National University, Canberra, ACT 2602, Australia

**DOI:** 10.21105/joss.00232

**Software**

- Review
- Repository
- Archive



## Summary

Libqcpp implements a variety of algorithms for Next-generation Sequencing (NGS) data quality control. These algorithms include:

- Sliding-window quality score trimming, using an algorithm based on Sickle (Joshi and Fass 2011).
- A combined adaptor removal and read merging algorithm for paired end reads that uses global pairwise alignment of reads. This algorithm is similar to AdapterRemoval (Lindgreen 2012).
- Cycle-wise summarisation of base quality scores, similar to FastQC (Andrews 2012)

Libqcpp allows simple composition of quality control pipelines that combine these features into a single unit. Application code can then simply read from a stream of sequence reads that have passed quality control measures. Optionally, parsing and quality control can occur in one or more background threads for efficiency. Reports detailing actions performed and summaries of results may be obtained in YAML format. Libqcpp includes `trimit`, a command line interface to these features for those not building their own applications.

Libqcpp uses the SeqAn library for sequence parsing and alignment (Döring et al. 2008), libyaml-cpp for YAML report generation, and Catch for unit testing. Documentation on API and command line usage is included, and available at https://qcpp.readthedocs.io/.

## References

Andrews, S. 2012. "FastQC A Quality Control Tool for High Throughput Sequence Data." http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Döring, Andreas, David Weese, Tobias Rausch, and Knut Reinert. 2008. "SeqAn an Efficient, Generic C++ Library for Sequence Analysis." *BMC Bioinformatics* 9: 11. doi:10.1186/1471-2105-9-11.

Joshi, N A, and J N Fass. 2011. *Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files* (version 1.33). https://github.com/najoshi/sickle.

Lindgreen, Stinus. 2012. "AdapterRemoval: Easy Cleaning of Next-Generation Sequencing Reads." *BMC Research Notes* 5: 337. doi:10.1186/1756-0500-5-337.

# Chapter 4

# kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity

This chapter presents a new computational method for estimating genetic distance from short read sequencing data. Specifically, kWIP extends Euclidean distance-based alignment-free sequence comparison methods with a novel weighting scheme. This entropy weighting reduces the noise introduced by sequencing error and variable coverage inherent in low-coverage population resequencing experiments. I both devised the novel metric, designed and implemented the software, and conducted validation experiments. Co-authors suggested mathematical improvements to the implementation of the metric, and helped design validation experiments.

This work was accepted in PLoS Computational Biology (2017; doi: 10.1371/journal.pcbi.1005727). The senior author authorises the inclusion of this manuscript in my thesis.

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# kWIP: The *k*-mer weighted inner product, a *de novo* estimator of genetic similarity

**Kevin D. Murray[1]*, Christfried Webers[2,3], Cheng Soon Ong[2,3], Justin Borevitz[1], Norman Warthmann[1]***

**1** Research School of Biology, The Australian National University, Canberra, Australia, **2** Data61, CSIRO, Canberra, Australia, **3** Research School of Computer Science, The Australian National University, Canberra, Australia

* kdmpapers@gmail.com (KDM); norman@warthmann.com (NW)



OPEN ACCESS

**Citation:** Murray KD, Webers C, Ong CS, Borevitz J, Warthmann N (2017) kWIP: The *k*-mer weighted inner product, a *de novo* estimator of genetic similarity. PLoS Comput Biol 13(9): e1005727. https://doi.org/10.1371/journal.pcbi.1005727

**Editor:** Andreas Prlic, UCSD, UNITED STATES

**Received:** October 4, 2016

**Accepted:** August 21, 2017

**Published:** September 5, 2017



**Data Availability Statement:** The software, kWIP, is publicly available at https://github.com/kdmurray91/kWIP. All simulation data was created by supplied reproducible workflows. All other analysis is based on published data that is publicly available and referenced.

**Funding:** This project was supported by the Australian Research Council Centre of Excellence in Plant Energy Biology (CE140100008) and by NICTA which was funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence

## Abstract

Modern genomics techniques generate overwhelming quantities of data. Extracting population genetic variation demands computationally efficient methods to determine genetic relatedness between individuals (or "samples") in an unbiased manner, preferably *de novo*. Rapid estimation of genetic relatedness directly from sequencing data has the potential to overcome reference genome bias, and to verify that individuals belong to the correct genetic lineage before conclusions are drawn using mislabelled, or misidentified samples. We present the *k*-mer Weighted Inner Product (`kWIP`), an assembly-, and alignment-free estimator of genetic similarity. `kWIP` combines a probabilistic data structure with a novel metric, the weighted inner product (WIP), to efficiently calculate pairwise similarity between sequencing runs from their *k*-mer counts. It produces a distance matrix, which can then be further analysed and visualised. Our method does not require prior knowledge of the underlying genomes and applications include establishing sample identity and detecting mix-up, non-obvious genomic variation, and population structure. We show that `kWIP` can reconstruct the true relatedness between samples from simulated populations. By re-analysing several published datasets we show that our results are consistent with marker-based analyses. `kWIP` is written in C++, licensed under the GNU GPL, and is available from https://github.com/kdmurray91/kwip.

This is a *PLOS Computational Biology* Software paper.

## Introduction

A major application of DNA sequencing is comparing the genetic make-up of samples with one another to either identify commonalities, and thus detect relatedness, or to leverage the differences to elucidate function. Initially, one seeks to confirm assumed genetic lineages and

Program. The research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government. KDM is supported by an Australian Government Research Training Program (RTP) Scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

replicates or to group samples into families, populations, and species. Estimating the genetic relatedness between a broad collection of samples must avoid bias and have minimal per sample cost.

Nowadays, the vast majority of studies in population genomics are performed using next generation sequencing (NGS) [1]. The methods commonly employed to analyse whole genome DNA sequencing data rely on two complementary concepts: the assembly of reference genomes and comparing samples to this reference by re-sequencing, read mapping, and variant calling. This approach, while functional in model organisms, is not ideal. Selecting the reference individual is mostly random, generating a reference genome assembly is time consuming and costly [2, 3], and analyses based on read alignment to a possibly inappropriate reference genome sequence are highly susceptible to bias [4, 5], to the point where large parts of the genomes are missed when sufficiently different or absent from the reference. Alignment-free methods for measuring genetic relatedness would help overcome this reference genome bias.

Another issue of concern is sample identification. A recent review [6] found that sample misidentification occurs at an alarming rate. With ever increasing sample numbers in (population) genetic projects, the issue of correct and consistent metadata arises on several levels: technical (mix-up) and biological (misidentification). Large field, and entire gene bank collections are being DNA-sequenced. With sample handling from the field through the laboratory to the sequence read files and eventual upload to data repositories, there is ample opportunity for mix-up and mislabelling of samples and files. This problem is exacerbated by the often highly collaborative nature of such undertakings. Some misidentifications, however, might be virtually undetectable without molecular genetic analysis, such as varying levels of ploidy, cryptic species, or sub-genomes in (compilo)species complexes [7]. Unfortunately, much of this hidden variation is easily overlooked by following aforementioned current best practices to calculate genome-wide genetic relatedness from short read sequencing data. Erroneous sample identification and/or underestimating the level of divergence has implications for downstream analysis choices, such as which samples and populations to use for a Genome Wide Association Study (GWAS); the missing heritability might then in fact be in the metadata.

The field of alignment-free sequence comparison aims to combat these difficulties by avoiding the process of sequence alignment. Approaches include decomposition into words, i.e., substrings of length k, commonly referred to as *k*-mers [8–11], sub-string or text processing algorithms [12–14], and information theoretic measures of sequence similarity or complexity [15]. While avoiding sequence alignment, some alignment-free sequence comparison tools still require prior knowledge of the underlying genome sequences, which precludes their use as a *de novo* tool. Recently, several algorithms enabling *de novo* comparisons have been published. These extensions all attempt to reconstruct phylogenetic relationships from sequencing reads. `Spaced` [13, 16] uses the Jensen-Shannon distance on spaced seeds (small *k*-mers a short distance from one another or with interspersed disregarded bases) to improve performance of phylogenetic reconstruction. `Cnidaria` [17] and `AAF` [18] use the Jaccard distance to reconstruct phylogenies, while `mash` [19] uses a MinHash approximation of Jaccard distance to the same effect.

One of the most established and studied alignment-free sequence comparison metrics is the $D_2$ statistic [8, 10]. It measures the difference between two sequences by the number of *k*-mer matches. First, all *k*-mers are counted in each sequence and recorded in a count vector. Then the difference between those vectors is measured. In the case of the original $D_2$ statistic, this is achieved by simply building the vector product. Several derivatives of the $D_2$ statistics, e.g., $D_2^*$, $D_2^S$, have been developed over the years [8, 20–23], which aim to improve accuracy by modelling and correlating observed versus expected *k*-mer frequencies. While these statistics

have been extended to Next Generation Sequence data [24] and successfully applied to metagenome comparisons [25], these $D_2$ statistic derivatives, such as $D_2^*$ and $D_2^S$, have the significant drawbacks of slow computational speed and the difficulties of defining the background models.

Here we present the *k*-mer Weighted Inner Product, a new metric to estimate genetic relatedness that introduces and combines two concepts to *k*-mer-based sequence comparison. Similar to the $D_2$ statistic(s), the similarity measure is an inner product of *k*-mer counts, but firstly, we no longer compare every *k*-mer, but rather hash all *k*-mers of a sample into a probabilistic data structure: a sketch [26]. The resulting sketches are, in effect, vectors of *k*-mer counts; importantly, the sketches for all samples have a constant size. Secondly, we introduce an information-theoretic weighting to elevate the relevant genetic signal above the noise. Pairwise similarity is then calculated by the inner product between *k*-mer counts, weighted by the information content derived from their frequencies across the population. Our procedure is implemented in a software tool (`kWIP`) that calculates our metric, the *k*-mer Weighted Inner Product, directly from sequencing reads. We show by simulations and by re-analysing published datasets, that `kWIP` can quickly, and accurately detect genetic relatedness between samples.

## Design and implementation

`kWIP` operates on files containing sequencing reads generated by common modern sequencing platforms (e.g., Illumina). First, `kWIP` utilises `khmer` [27, 28] to count overlapping words of length *k* (*k*-mers) into a probabilistic data structure, a sketch, for each sample. In order to establish the weights `kWIP` then counts presence/absence of each *k*-mer across all sample sketches and records this population occurrence frequency in a frequency sketch (*F*). We calculate similarity (*K*) as the inner product between each pair of sample sketches, weighted by the Shannon entropy (*H*) of the respective frequency (*F*). The concept is illustrated in Fig 1.

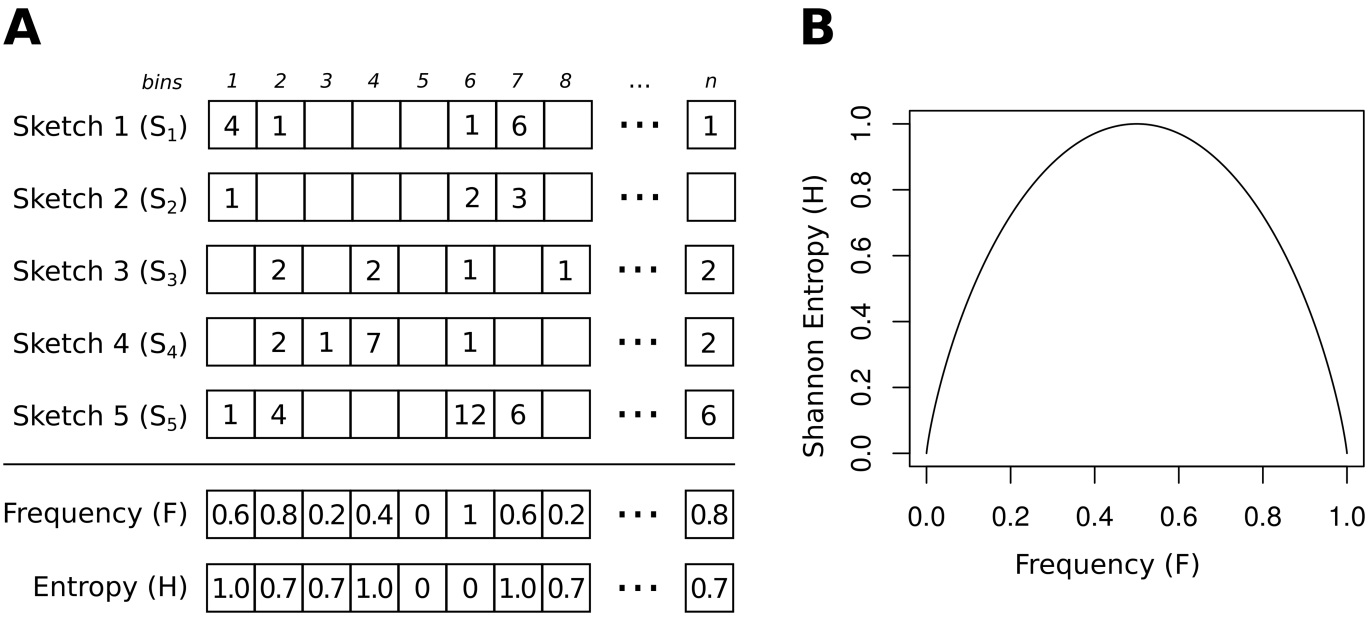



**Fig 1. Overview of the weighted inner product metric as implemented in `kWIP`.** (A) *k*-mers are counted into sketches (using `khmer` [28]). Columns represent the "bins" in each sketch. The frequencies of non-zero counts across a set of sketches is computed, forming the population frequency sketch (denoted *F*). We calculate Shannon entropy of this frequency sketch as the weight vector for the WIP metric (denoted *H*, see Eq 2). (B) Illustration of Shannon Entropy as used in `kWIP`: the relationship between the population frequency (*F*) and the weight (*H*).

https://doi.org/10.1371/journal.pcbi.1005727.g001

## *k*-mer counting

For each sample, `kWIP` uses `khmer` to decompose sequencing reads into overlapping words of some fixed length k, e.g., 20. The value of a reversible hash function is computed for each *k*-mer. *k*-mers are canonicalised by using the lexicographically smaller of a *k*-mer and its reverse complement. *k*-mers are counted using one sketch per sample. These sketches are vectors with prime number length, typically several billion elements in size (denoted $S_i$ for sample *i*). The elements of these sketches are referred to as bins (indexed by *b*, e.g. $S_{i_b}$), and can store values between 0 and 255 (integer overflow is prevented). To count a *k*-mer, the *b*-th bin of the sketch ($S_{i_b}$) is incremented, where *b* is the hash value of the *k*-mer modulo the (prime) length of the sketch. For most use cases, *k*-mers between 19 and 21 bases long should achieve a good balance between specificity and sensitivity across genomes and genomic regions [29]. Note that the possible number of *k*-mers ($4^k$) is much larger than the length of a sketch. Therefore, aliasing (or "collisions") between *k*-mers can occur, but in practice can be avoided with appropriate parameter selection [27]. It is worth noting that aliasing can only increase similarity between any two samples and should occur uniformly across all sample pairs.

## Weighting and similarity estimation

Genetic similarity is estimated by calculating the inner product between each pair of sample sketches ($S_i$, $S_j$), weighted by the informational content of each bin. The population frequency sketch (*F*) contains the frequency of occurrence for each bin, calculated as the proportion of samples with a non-zero count for each bin. We calculate a weight vector (*H*) of these occurrence frequencies using Shannon entropy as per Eq (1). In the Weighted Inner Product (WIP) metric (or kernel), pairwise similarities are then calculated as the inner product over every pair of sample sketches, weighted by *H* as per Eq (2). The unweighted Inner Product (IP) metric is simply the inner product between the two sketch vectors, $S_i^T S_j$, without weighting. This produces a matrix of pairwise inner products *K*, commonly referred to as a kernel matrix. The kernel matrix is then normalised using the Euclidean norm Eq (3), and converted to distances using the "kernel trick" [30] as per Eq (4). To ensure distance matrices are Euclidean, `kWIP` confirms that the resulting kernel matrix is positive semi-definite by checking that all eigenvalues are non-negative using the Eigen3 library [31].

The distances kWIP produces are relative within the set of samples being compared. This is because the weight vector (*H*) is specific to the set of samples and the similarity estimates are normalised to account for varying sequencing coverage. In other words, the kWIP distance for a given pair of samples will depend on the set of samples within which they are analysed.

$$H = -(F\log_2(F) + (1-F)\log_2(1-F)) \tag{1}$$

$$K_{ij} = \sum_{b=1}^{n} S_{i_b} S_{j_b} H_b \tag{2}$$

$$K'_{ij} = \frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}} \tag{3}$$

$$D_{ij} = \sqrt{K'_{ii} + K'_{jj} - 2K'_{ij}} \tag{4}$$

## Implementation

Pairwise calculation of genetic distances from *k*-mer count files with both the WIP and IP metrics is implemented in C++ as `kWIP`. `kWIP` is licensed under the GNU GPL, and source code and pre-compiled executables are available from https://github.com/kdmurray91/kwip. Documentation and tutorials are available from https://kwip.readthedocs.io. To use `kWIP`, one first counts *k*-mers present in each sample using `khmer`'s `load-into-counting.py` script [28]. `kWIP` will then estimate similarity from these counts, producing a normalised Euclidean distance matrix and, optionally, the corresponding similarity matrix (kernel matrix). `kWIP` parallelises pairwise similarity calculations across cores of a multi-threaded computer to ensure fast operation.

# Results

We show that `kWIP` is able to accurately determine genetic relatedness in many scenarios. Using a simulated population re-sequencing experiment, we quantify how the population frequency-based weighting applied by `kWIP` improves accuracy, that is the correlation with the known truth, when compared to existing approaches, `mash` [19], and the unweighted metric, IP. We recover known technical and biological relationships between sequencing runs of the 3000 Rice Genomes project [32, 33]. We show that `kWIP`'s estimate of genetic relationships between *Chlamydomonas* samples is nearly identical to results obtained by a more traditional, SNP-based analysis employing read mapping and variant calling against a reference genome with the same sequencing data [34]. By analysing a dataset on root-associated microbiomes [35], we show that our approach of sample clustering by `kWIP` can be extended to clustering of metagenome samples.

## Quantification of `kWIP` performance

We quantified the performance of `kWIP` with simulated population sequencing data. We compare our novel metric, the weighted inner product (WIP), to the unweighted inner product (IP), which we consider equivalent to the $D_2$ statistic, and to `mash` [19]. We simulated 20 populations of 12 individuals with 1 MBp genomes and analysed each with `kWIP` and `mash` for *k*-mers of k = 20. A summary of the results of these 20 replicate analyses with each of the metrics is shown in Fig 2.

Unsurprisingly, for all metrics the accuracy, that is the rank correlation (Spearman's $\rho$) to known truth, decreases with decreasing genome coverage, i.e., average sample sequencing depth (Fig 2A), as well as with decreasing average number of nucleotide differences per site, $\pi$ (Fig 2B).

Importantly, at low coverages, the weighted metric ("WIP") performs better than the unweighted ("IP") (Fig 2A). Above a certain coverage, in the case of our simulations above about 30-fold, the performances of the WIP and IP metrics converge. At a constant genome coverage, the improvement in accuracy of the WIP metric relative to the IP metric increases as mean pairwise genetic variation decreases (Fig 2B). While the accuracy of the IP metric decreases markedly below an average number of nucleotide differences per site ($\pi$) of approximately 0.01, the WIP metric does not show such decrease.

In order to compare the performance of `kWIP` relative to Mash [19] we conducted two analyses with `mash`: one with abundance filtering enabled to remove singleton *k*-mers ("Mash AF") and one without ("Mash"). Within the scope of our simulations `kWIP` yields more accurate results than `mash` when sequencing coverage and/or sequence divergence is low; a typical scenario in large-scale, population genetic analyses within species. Through the entire range of simulation parameters, `kWIP` never yields results less accurate than `mash`, irrespective of

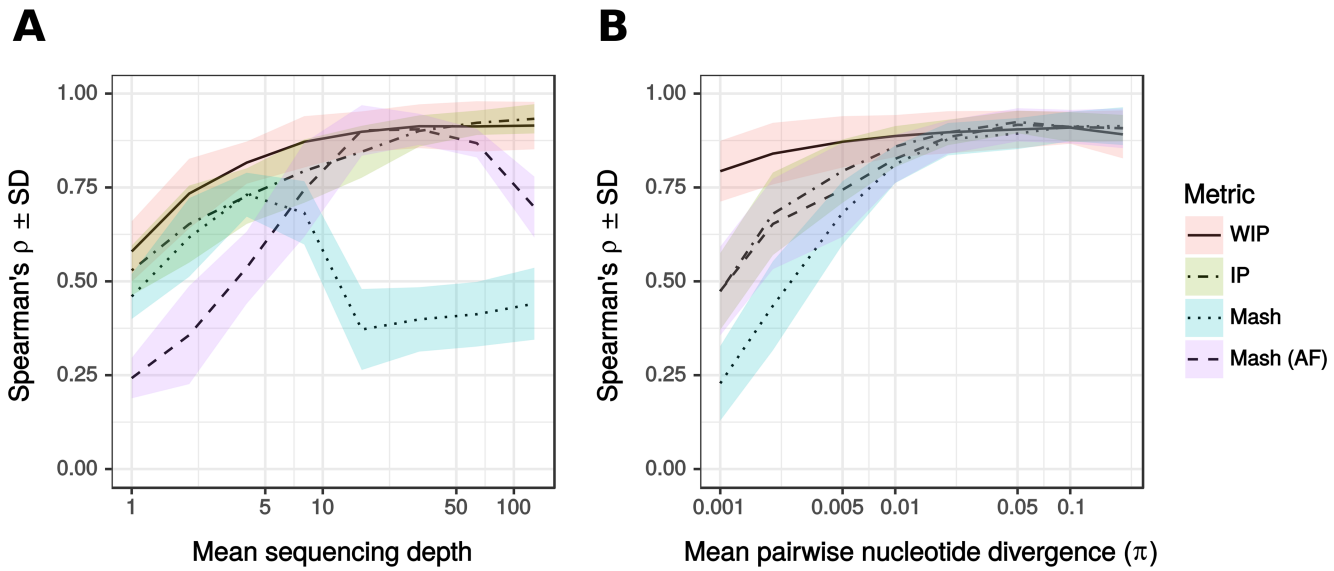



**Fig 2. The effect of (A) mean sequencing depth (genome coverage) and (B) average number of nucleotide differences per site ($\pi$) on accuracy of genetic similarity estimates in simulations.** We plot mean ± standard deviation of Spearman's $\rho$ comparing each metric to known truth across 20 replicate runs. (A) Mean sequencing depth varies while average number of nucleotide differences per site ($\pi$) is constant at 0.005. `kWIP`: At low to moderate mean sequencing depth (<30x) weighting increases accuracy. The weighted metric ("WIP") obtains near-optimal accuracy already at 10x and hence much earlier than the unweighted metric "IP"). There is no noticeable decrease in accuracy with increasing coverage. `mash`: regardless of error correction, `mash` performs less well than WIP. `mash` shows accuracy maxima at 4x coverage without ("Mash") and at 16x coverage with abundance filter ("Mash (AF)"), at which point Mash (AF) performs almost as well as WIP. The accuracy of `mash` decreases dramatically when coverage is further increased. (B) Genome coverage is kept constant at 8x and average number of nucleotide differences per site ($\pi$) varies. While all metrics perform equally at a ($\pi$) of 1 in 100 (0.01), the performance of IP, Mash and Mash (AF) decreases rapidly as ($\pi$) between samples decreases. This does not occur for the weighted metric (WIP).

https://doi.org/10.1371/journal.pcbi.1005727.g002

abundance filtering (Fig 2). It is interesting to note that `mash` appears to exhibit characteristic accuracy maxima, and accuracy decreases dramatically when mean sequencing depth is further increased. In addition, abundance filtering seems to have a strong, genome coverage-dependent effect on the accuracy of `mash` (Fig 2A). With the chosen parameter settings, `mash` runs much faster than `kWIP` (about 10-fold faster; see performance comparisons in Table 1).

In analyses with `kWIP` we find that the coefficient of variation between the number of sequencing reads per sample matters. For samples with much lower mean sequencing depth than the average, `kWIP` has difficulty to accurately determine its relatedness to other samples.

**Table 1. Computational performance of `kWIP`.**

| Dataset | Dataset Size | | | Distance Calculation Time (s) | | |
|---|---|---|---|---|---|---|
| | Samples | Reads | *k*-mers | Mash | WIP | IP |
| Simulation (8x) | 36 | 7.9e4 ± 2.4e3 | 1.3e6 ± 1.3e5 | 6 ± 1 | 45 ± 3 | 40 ± 4 |
| Simulation (32x) | 36 | 3.2e5 ± 1.0e4 | 2.3e6 ± 3.8e5 | 5 ± 1 | 53 ± 3 | 46 ± 5 |
| Rice Replicates | 96 | 9.7e6 ± 1.5e6 | 1.8e8 ± 1.5e7 | - | 2241 ± 139 | 1892 ± 286 |
| Chlamydomonas | 20 | 2.0e8 ± 2.4e7 | 1.4e8 ± 1.4e7 | - | 127 | 194 |

Measurements of calculation time are in wall-clock seconds on a 16-core, 64GB GNU/Linux server. Figures are means ± standard deviations. For simulations, these are over the 20 replicate runs performed. For rice replicate clustering, these are over 10 of the 100 independent sets of 96 rice samples. For Chlamydomonas, times are for the full dataset. The sketch sizes used were $10^9$ bins for `kWIP`/khmer, and $10^4$ for Mash. Note that *k*-mers refers to the number of distinct *k*-mers as estimated by khmer.

https://doi.org/10.1371/journal.pcbi.1005727.t001

We therefore advise to exclude such samples from `kWIP` analyses or sub-sample reads from the remainder, if the dataset allows. `khmer` provides procedures for "digital normalisation", which can be used upstream of `kWIP` to that effect [36]. Our simulations suggest that variations in genome coverage between samples will also affect the results obtained with `mash`.

## Replicate clustering

`kWIP` can efficiently verify replicates. Fig 3A and 3B show a representative example of replicate clustering. The weighted metric (WIP) is able to accurately cluster replicates (Fig 3A), whereas the unweighted metric (IP) makes mistakes, as highlighted in red in Fig 3B. We quantified this difference in performance and Fig 3C shows the distribution of rank correlation coefficients between distances obtained with the WIP and IP metrics and the expected clustering patterns for 100 sets of 96 sequencing runs. The WIP metric outperforms the IP metric, having a significant higher mean correlation (paired Student's T test, $n = 84$, $t = 9.63$, $df = 83$, $p = 3.6 \times 10^{-15}$).

## Population structure

Flowers, et al. [34] sequenced 20 strains of *Chlamydomonas reinhardtii*; laboratory strains and wild accessions sourced from across the continental USA. By alignment- and SNP-based analysis, they find significant population structure that is mostly explained by geography [34]. In Fig 4B we display the published genetic relationships as a principal component analysis (PCA) of SNP genotypes calculated with SNPRelate [37] exactly as presented by the authors [34]. PC1 separates the laboratory strains (and one western sample) from both eastern and western samples with further structure among wild *Chlamydomonas* accession collected in western, southeastern and northeastern USA. In Fig 4A we plot the relatedness between the strains as revealed directly from the raw sequencing reads with `kWIP`. We note that the results are highly similar; the rank correlation between `kWIP` distances and genome average identity-by-state (calculated with SNPRelate [37]) is 0.95.

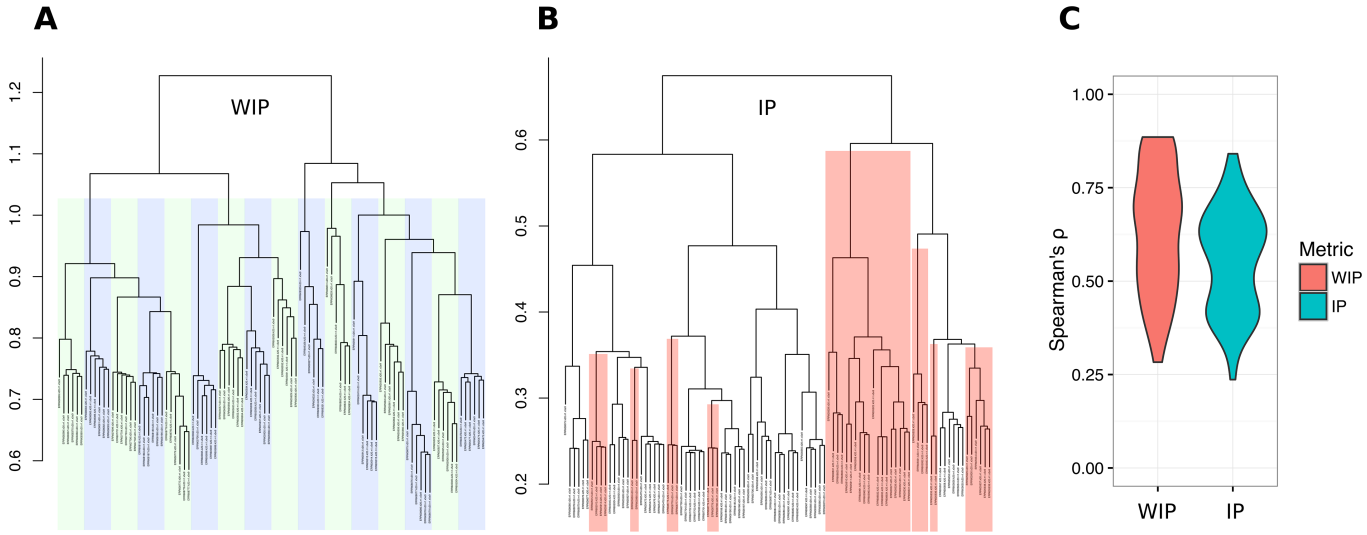



**Fig 3. Weighting improves the accuracy of replicate clustering.** (A) and (B) show a representative example, demonstrating that (A) the weighted metric (WIP) correctly clusters all sets of 6 replicate runs into their respective samples (indicated by blue and green bars) while (B) the unweighted metric (IP) fails to cluster several replicates correctly (indicated by red highlighting). (C) rank correlation coefficients to expected relationships over 100 sets of 96 rice runs for the WIP and IP metrics. The Weighted metric tends to cluster the replicates better.

https://doi.org/10.1371/journal.pcbi.1005727.g003

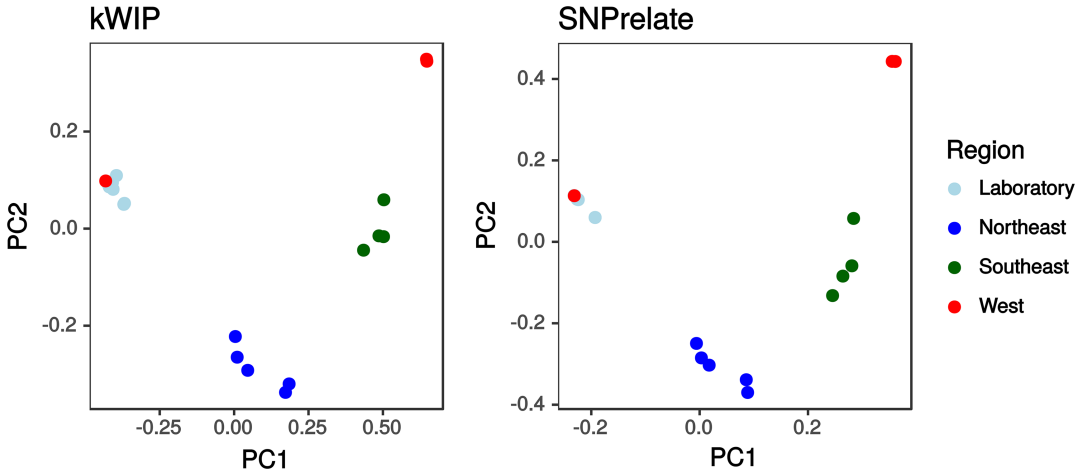



**Fig 4. Genetic relatedness between *Chlamydomonas reinhardtii* strains based on sequencing data from [34].** SNPrelate [37] was used to compute the PCA decomposition directly from SNP genotypes provided by the authors. This replicates the analysis of [34] and is displayed on the right. On the left, we show the results of MDS performed on the distance matrix obtained with `kWIP`.

https://doi.org/10.1371/journal.pcbi.1005727.g004

Each of the 20 strains had been sequenced to a depth of roughly 200-fold genome coverage [34]. By systematically sub-sampling this dataset we investigated the effect of coverage on the accuracy of `kWIP`'s similarity estimation. We find that with decreasing coverage the accuracy of the relationship estimations decreases (Fig 5A). We illustrate this decay by PCA plots of estimated genetic relatedness at varying coverages (Fig 5B). We note that the performance of `kWIP` to determine similarity is very good even at low coverages. A two-fold genome coverage is enough to detect the major splits in this dataset (Laboratory vs West vs East).

## Metagenome relatedness

Edwards, *et al.* [35] sequenced 16S rDNA amplicons from rice root-associated microbiomes and find stratification of samples by rhizo-compartment, cultivation site, and cultivation practice. Analysing their raw sequencing data with `kWIP`, we detect highly similar stratification between microbial communities. An example is shown in Fig 6. We observe a gradient of samples from within the root, through the root-soil interface into soil, and separation by cultivation site. This replicates the separation of samples by rhizo-compartment and cultivation site published by Edwards, et al. [35], shown in Fig 6.

# Discussion

The *k*-mer Weighted Inner Product (`kWIP`) estimates genetic distances between samples within a population of samples directly from next generation sequencing data. `kWIP` does not require a reference genome sequence and is able to estimate the genetic distances between samples with less data than is typically used to call SNPs against a reference. As a *k*-mer-based method, `kWIP` is sequencing protocol and platform agnostic, allowing use into the future.

`kWIP` uses a new metric, the weighted inner product (WIP), which aims to reduce the effect of technical and biological noise and elevate the relevant genetic signal by weighting *k*-mer counts by their informational entropy across the analysis set. This weighting has the effect of down-weighting *k*-mers that are either highly abundant or present in very few samples. Those *k*-mers are typically uninformative, because they are either common, fixed, repetitive,

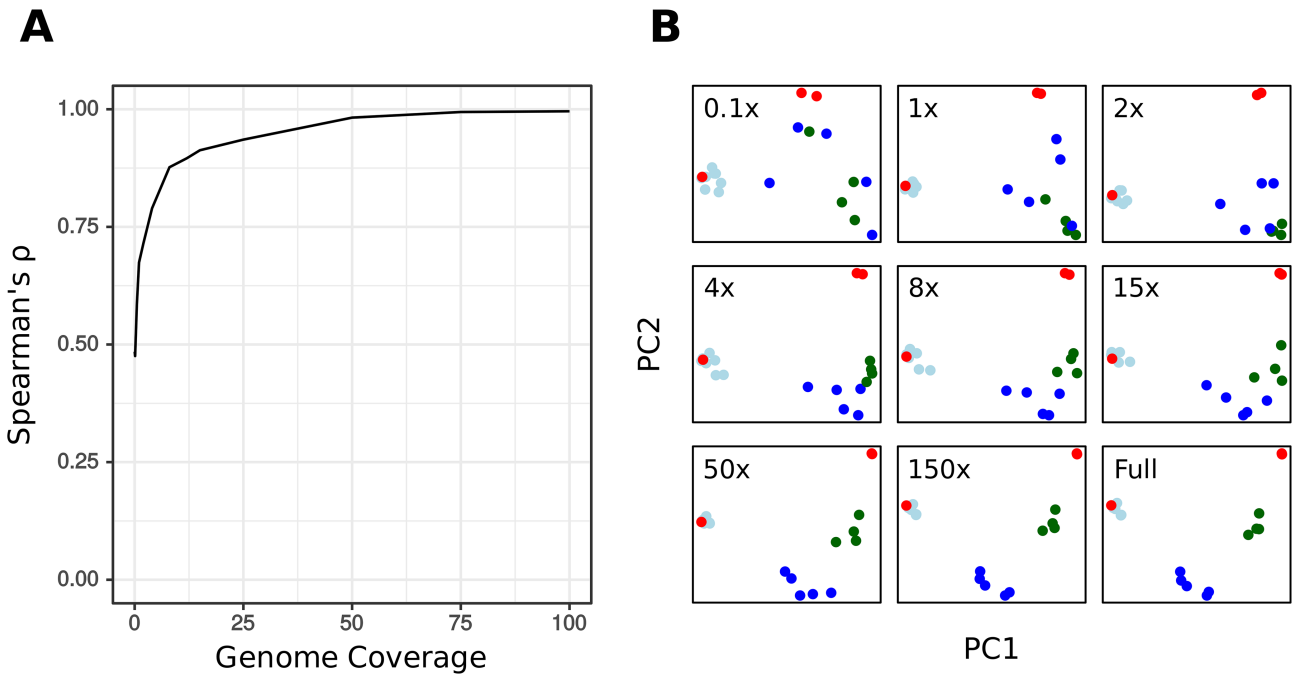



**Fig 5. The effect of mean sequencing depth (genome coverage) on `kWIP`'s estimate of genetic relatedness between samples of *Chlamydomonas reinhardtii* (data from [34]).** (A) Spearman's rank correlation between sub-sampled datasets and the full dataset across a range of subset average genome coverages. (B) PCA plots of relatedness obtained using `kWIP` on selected sub-sampled datasets. "full" refers to the entire dataset (i.e., Fig 4), while "0.1x" refers to a sub-sampled dataset with average mean sequencing depth of 0.1 over the *C. reinhardtii* genome (likewise for 1x, 2x, and so on).

https://doi.org/10.1371/journal.pcbi.1005727.g005

invariable, or rare, or erroneous. By using Shannon entropy, the weights of common and infrequent *k*-mers are assigned lower, but non-zero weights, allowing them to contribute to the signal.

Euclidean distances are then calculated from these weighted inner products and `kWIP` outputs a matrix of pairwise distances between samples, which are easily visualised and may be

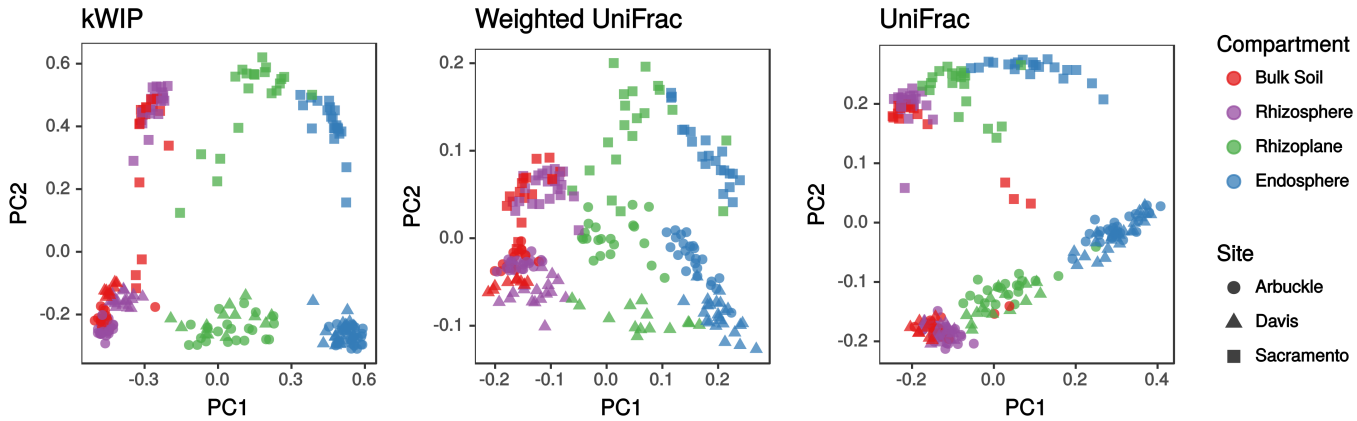



**Fig 6. Estimation of similarity between metagenome samples.** We used `kWIP` to examine 16S rDNA amplicon sequencing data of Edwards, *et al.* [35] and compare our `kWIP` result ("kWIP") with the results as presented by Edwards, *et al.* ("Weighted UniFrac" and "UniFrac"). We find that `kWIP` replicates their observations of stratification of root-associated microbiomes by rhizo-compartment (PC1) and experiment site (PC2).

https://doi.org/10.1371/journal.pcbi.1005727.g006

used for sample classification and to cluster samples into groups. These distance matrices are amenable to quantitative comparison of genetic distance to geographic or environmental distances, for example using mantel tests or generalised dissimilarity modelling. We show high concordance between PCAs obtained using SNP data and those using `kWIP`. It is possible that population genetic statistics, including $F_{ST}$, could be recovered using `kWIP` via a genealogical interpretation of PCA, as is proposed and shown possible for SNP datasets [38].

We have demonstrated the applicability and effectiveness of `kWIP` using simulations and several published datasets. Through simulations, we quantify how the novel weighting improves accuracy over both the unweighted inner product and Mash, specifically in cases where genetic differentiation or sequencing depth is low (Fig 2). With data from the 3000 rice genome dataset [33], we reconstruct known relationships between samples and sequencing runs, such as membership of samples to major genetic groups of *Oryza sativa*, and the correct clustering of replicates (Fig 3).

From sequencing reads of a population re-sequencing experiment in *Chlamydomonas* [34] we precisely recreate their visualisation of population relatedness (Fig 4). This dataset suited for comparison because Flowers, *et al.*, had based their analysis not only on variants recovered by read alignment to the reference genome, but attempted to recover and use additional variation by assembling leftover reads that did not match the reference into contigs and calling additional variants between these contigs. This approach, while reducing reference-genome bias, required extensive sequencing depth to enable de-novo assembly; the authors chose around 200-fold coverage, which in turn enabled us to assess `kWIP`'s performance at various sequencing depths (Fig 5).

Efficient characterisation of complex metagenome samples has traditionally relied on methods of reduced representation. We show that `kWIP` is able to detect structure between microbial communities based on 16S rDNA amplicon sequencing data, at least as well as current practice (Fig 6). It should be possible to apply `kWIP` to random shotgun sequencing data from such samples. Also, estimates of complexity and diversity within and between metagenomes are currently mostly gene based, but could also be made efficiently at the *k*-mer-level leveraging sketched data structures.

The key innovation of `kWIP` is the combination of a fixed-sized, probabilistic data structure (sketch) for counting *k*-mers with an entropy-weighted inner product as a measure of similarity between samples. By virtue of their fixed size, sketches enable rapid arithmetic operations on *k*-mer counts. Sketches enable `kWIP` to rapidly aggregate across a populations to derive weights, and to efficiently compute the inner products. These benefits outweigh the possibility of collisions between *k*-mers, which in any case have been observed to be rare [27] given appropriate sketch size. Sketching data structures are commonly used for *k*-mer counting (for example Count-Min Sketches [27, 28], and Bloom Filters [39]), but have not been widely adopted in alignment-free sequence comparison.

Weighting of inner products between sketches allows us to account for non-uniform information content of each *k*-mer. `kWIP` weights by Shannon entropy of presence/absence frequency across a population. This provides an assumption-free estimate of the information content of each *k*-mer. By down-weighting both rare *k*-mers introduced by rare variants or sequencing errors, as well as *k*-mers present in most or all samples, we are reducing the contribution of *k*-mers that carry less information. It is possible that other weighting functions that assume various population parameters could provide a more faithful estimate of the information content of each *k*-mer. The application of word-specific weighting has precedence in text processing, where it has been used to account for varying importance of words in a document [40]. However, because we intend `kWIP` to be used in situations where such

parameters are either unavailable or potentially inaccurate, we prefer that our weighting is free of assumptions.

An inner product between *k*-mer counts has long been used to detect and measure sequence similarity, and is referred to as the $D_2$ statistic. There have been many derivatives of the $D_2$ statistic that seek to enhance its accuracy in recreating evolutionary histories (e.g., $D_2^S$, and $D_2^*$ [20–22]). `kWIP` does not attempt to re-create evolutionary histories, but rather estimates the similarity of genetic material as it exists today. This is sufficient and even desirable for many of `kWIP`'s intended uses. When validating experimental metadata, one seeks to establish whether similarity between sequencing runs matches expectations. Particularly for metagenome samples, where variation can be in both abundance and type of organisms, estimating present variation between sample genome sequences is of importance, separate to how this variation came to be.

`kWIP` estimates genetic similarity between sequencing runs. Because `kWIP` operates reference- and alignment-free, all genetic material present in the sample, the "hologenome", will contribute to the analysis. However, we note that *k*-mers that are considered undesirable and chosen to be excluded from the analysis could easily be masked, for example by setting their weight in the weight vector to zero.

Because `kWIP` weights *k*-mers, and hence genome content, based on their frequency in the population being analysed, these weights change when the population changes. This allows for iterative workflows: in a first, all inclusive step the large groupings and outliers are detected; subsequently, subgroups can be analysed with increased resolution.

`kWIP` is purposefully designed to operate free of assumptions or prior knowledge. It is comparing data as presented in the sequencing reads without attempting to reconstruct or approximate the underlying genomes. One could think of several ways of incorporating additional knowledge, which may improve `kWIP`'s power to determine relatedness between underlying genomes. One could, for example, apply smoothing to the *k*-mer counts, with the goal of differentiating between *k*-mers that are genuinely not in the genomes of a sample and those that were not observed due to low coverage and/or stochastic sampling; smoothing is used in natural language modelling [41].

It is possible that alternative distance functions (e.g., Manhattan distance) over weighted sketches could improve the performance of `kWIP`, which currently uses Euclidean distance. Distance measures defined on presence/absence of items, such as the Jaccard index or the Jaccard index-based measures used by AAF [18] and `mash` [19], could also be calculated from our sketches. It may further prove valuable to explore spaced seeds [13, 16], or alternative metrics including those considering inexact matches [42, 43].

Methods that enable rapid verification of genetic resources, such as stock centre accessions or cell lines, prevent expensive and possibly catastrophic mis-identifications. Such classification tasks only require comparison with a set of reference samples rather than computing distances between all samples. Inner product kernels have been used to classify protein sequences [43, 44] and `kWIP` could be adapted to sample classification with tree-like structures of kernels [42] or sketches [45, 46].

Estimating the genetic relatedness between a broad collection of natural accessions provides a basis for ecological or functional studies and should be a first step towards solutions in breeding and conservation. In most population level experiments, technical sources of error are dwarfed by the error from insufficient sampling [47]. This is especially true when rare or cryptic lineages are present, and in conditions of non-random mating where population structure is substantial. Such population level noise can only be overcome by broad studies with large numbers of samples, ideally by also merging experiments [48]. When individuals from real-

world populations are collected, or collated, there is normally non-uniform genetic relatedness. Initially, one seeks to group samples into more closely related families or more distantly related populations, to then develop sets for further detailed studies. Genetic outliers can represent mis-identifications and cryptic species and should be detected and excluded. *De novo* sample groupings based on whole genome relatedness also inform the selection of suitable reference individuals and/or building the necessary reference genome sequences. The initial characterisation process must avoid biases and have minimal per sample cost. The use of `kWIP` allows one to base the analysis of diversity among samples on low coverage, whole-genome sequence data and thus facilitates large, balanced study designs. More broadly, experiments are condemned to be inconclusive and irreproducible if samples are somehow mislabelled or misidentified. An initial step in all analyses of genetic or functional variation must involve the verification of sample identity [6]. This preliminary analysis should preferably use whole-genome sequence data, be *de novo*, unbiased, and agnostic to sequencing protocol and technology. `kWIP` is an efficient implementation of such a tool.

## Availability and future directions

`kWIP` is implemented in C++ and licensed under the GNU GPL. Source code and pre-compiled executables are available from https://github.com/kdmurray91/kwip. Documentation and tutorials are available from https://kwip.readthedocs.io. Docker images, Snakemake workflows and Jupyter notebooks used to perform all analyses presented here are available online at https://github.com/kdmurray91/kwip-experiments; the respective software versions are noted within the repository. When given a population of samples, `kWIP` performs all pairwise comparisons, which scales quadratically with regards to the number of samples ($\mathcal{O}(n^2)$), but parallelises pairwise similarity calculations across cores of a multi-threaded computer to ensure fast operation. Analyses of very large data sets, i.e., beyond 10,000s of samples, will benefit from further optimisation to the implementation of `kWIP`, including parallelisation across distributed memory systems with MPI. For each pairwise comparison, the two sketches and the weight vector must fit in main memory. This limits the size of the sketches and the number of pairwise comparisons that will run efficiently in parallel on a given node.

## Materials and methods

We demonstrate `kWIP`'s performance with both real and simulated datasets. With simulations we quantify the performance of `kWIP`. To demonstrate the utility of `kWIP` in real-world, low-coverage, large-scale population genomics datasets, we analyse data from the 3000 Rice Genomes Project [32, 33]. To show that `kWIP` estimates genetic similarity as well as current best practice SNP-based methods, we re-analysed a population genomics study on 20 strains of *Chlamydomonas reinhardtii* [34] with `kWIP` and compare our result to the published results. Lastly, using data from a study on root-associated microbiomes of rice [35], we show that `kWIP` is able to separate microbial communities from 16S rDNA amplicon data at least as well as current best-practice methods in metagenomics.

We provide all information necessary to reproduce our work: the `kWIP` analyses performed here are implemented in Snakemake workflows [49], which describe all steps and software parameters; random seeds have been fixed where necessary. All downstream analyses are available as Jupyter notebooks [50, 51]. Both the Snakemake workflows and Jupyter notebooks are available online at https://github.com/kdmurray91/kwip-experiments; the respective software versions are noted within this repository.

## Simulations

We simulated several datasets to empirically quantify the performance of `kWIP`. Twenty populations with 12 individuals each were simulated using `scrm` [52]. Branch lengths within each population were normalised such that the mean pairwise genetic distance ($\pi$) was equal. Branch lengths were then scaled over a range of $\pi$ (between 0.001 and 0.2) to test the effect of mean pairwise genetic distance on accuracy. Genome sequences of 1 Mbp genomes were simulated with DAWG2 [53] and from those short read data for three replicate sequencing runs per individual were generated at various mean coverages (between 1- and 128-fold) using Mason2 [54]. We attempted to emulate the reality of sequencing experiments by introducing random variation in read numbers between replicate runs (coefficient of variation of 0.3). These simulated sequencing runs were then used to estimate genetic similarity with `kWIP` and `mash` [19]. For analysis with `kWIP` we used `khmer` to hash *k*-mers of length 20 into sketches with $10^7$ bins. We estimated genetic similarity with `kWIP`, using the weighted ("WIP") and unweighted ("IP") metrics. On the same data we performed two analyses with Mash, counting 20-mers into sketches of size $10^4$. For one analysis, we invoked the abundance filter within `mash sketch` such that only *k*-mers observed at least twice were considered ("Mash (AF)"), whereas the other analysis considered all *k*-mers regardless of abundance ("Mash").

The performance of our metrics was measured relative to the true pairwise distances between the simulated samples. The true distance matrix between samples was calculated from the simulated, aligned sample genomes (which DAWG2 produces) with `scikit-bio`. Sample-wise distances were replicated three times to allow comparison to the distances obtained from the three simulated sequencing runs. Performance was calculated as Spearman's rank correlation ($\rho$) between all pairwise distances using `scipy` [55].

## Datasets

With several published datasets we demonstrate the performance and utility of `kWIP` in real-world scenarios. In all cases, sequence data files for sequencing runs were obtained from the NCBI Short Read Archive using `sra-py` [56]. Reads were extracted using the SRA toolkit to FASTQ files. Low base quality regions were removed using sickle [57] in single-end mode. Counting of *k*-mers into count files (sketches) was performed using the `load-into-counting.py` script of `khmer`. Genetic similarity was estimated using `kWIP`, with the WIP and IP metrics.

To assess how well `kWIP` recovers replicate samples and known sample hierarchies at low sequencing coverage, we turned to publicly available sequence data from the 3000 Rice Genomes project [32, 33]. Samples of the 3000 Rice Genomes project had been sequenced on the Illumina HiSeq2000 platform with technical replicates of individual sequencing libraries split between 6 or more sequencing lanes [32, 33]. Furthermore, there is a rather strong subdivision of rice (*Oryza sativa*) into subgroups. We compiled 100 sets of 96 runs, i.e., for each set we chose 16 samples with 6 replicate runs. We ensured that 8 samples each were described by [32] as belonging to the Indica and Japonica subgroups of *O. sativa*. We estimated the genetic similarity between runs in each of these 100 sets with `kWIP`. The true distances between the different runs in the 3000 rice datasets are not known, but a topology and sample hierarchy can be inferred from the metadata. We hence assessed the performance of `kWIP` in accurately clustering replicates and recovering population structure against a mock distance matrix that reflects the expected topology. We created a distance matrix in which each run had a distance of zero to itself, a distance of 1 to each of its technical replicates (i.e., the other sequencing runs belonging to the same sample), a distance of 2 to each run from other samples in the same rice group (Indica or Japonica), and a distance of 4 to each run from a sample belonging to the

respective other rice group. We then used `scipy` to calculate Spearman's rank correlation between this mock matrix and each distance matrix obtained from real data using `kWIP`. A paired Student's t-test was performed between the estimates of relatedness from the WIP and IP metrics with the `t.test` function in R. We used hierarchical clustering to visualise these relationships, performed in R with the `hclust` function.

We use whole genome sequencing data on 20 strains of *Chlamydomonas reinhardtii* [34] to demonstrate that `kWIP` is able to detect population structure in a real-world dataset and to examine the effect of sample sequencing depth (coverage) on accuracy of `kWIP`. Genetic relatedness between the 20 *Chlamydomonas reinhardtii* samples from this study was estimated with `kWIP` using the WIP metric. Classic Multi-dimensional Scaling (MDS) of the `kWIP` distance matrix was performed using the `cmdscale` function in R. For Euclidean distance matrices, MDS is equivalent to PCA [58]. We compare our MDS results with the principal component analysis (PCA) decomposition of SNP genotypes calculated with function `snpgdsPCA` in SNPrelate [37], working from a VCF file provided by Flowers *et al.* [34]. From the aforementioned SNP data we calculated genome-wide average identity-by-state (IBS) with the `snpgdsIBS` function in SNPrelate [37]. Rank correlation between `kWIP` distances and 1-IBS was calculated with function `cor` in R [59].

We examined the effect of mean sequencing depth (coverage) on the accuracy of `kWIP` by random sub-sampling from the sequencing data of each sample. We sub-sampled to coverages of between 0.01- and 200-fold average genome coverage (0.01, 0.1, 0.5, 1, 2, 4, 8, 12, 15, 25, 50, 75, 100, 150, 200x) across samples using the `sample` command of `seqtk` [60]. We attempted to preserve the coefficient of variation in read numbers that existed in the original dataset (0.12) by sampling a random number of reads from the appropriate normal distribution. Spearman's rank correlation ($\rho$) was used to compare pairwise distances calculated at each sub-sampled coverage to those from the original dataset with function `cor` in R [59].

To demonstrate that `kWIP` can determine the relatedness of samples in a typical metagenomic dataset, we used next generation sequencing data from a study on rice root associated microbiomes [35] representing 16S rDNA amplicons from soil and root samples. Relatedness between samples was estimated using `kWIP` with the WIP metric, and MDS was performed as above.

## Acknowledgments

We thank Sylvain Forêt, Teresa Neeman, Conrad Burden, Gavin Huttley, Ben Kaehler, Cameron Jack and Fengzhu Sun for comments and advice on the metrics, algorithms, and experiments reported here. We thank Luisa Teasdale for comments on earlier versions of this manuscript. We thank Joseph Edwards and Johnathan Flowers for providing additional advice on and results from their datasets.

## Author Contributions

**Conceptualization:** Kevin D. Murray, Norman Warthmann.

**Data curation:** Kevin D. Murray.

**Formal analysis:** Kevin D. Murray, Christfried Webers, Cheng Soon Ong, Norman Warthmann.

**Funding acquisition:** Justin Borevitz.

**Investigation:** Kevin D. Murray, Norman Warthmann.

**Methodology:** Kevin D. Murray, Christfried Webers, Cheng Soon Ong.

**Project administration:** Norman Warthmann.

**Software:** Kevin D. Murray.

**Supervision:** Justin Borevitz, Norman Warthmann.

**Validation:** Kevin D. Murray, Norman Warthmann.

**Visualization:** Kevin D. Murray.

**Writing – original draft:** Kevin D. Murray, Christfried Webers, Cheng Soon Ong, Justin Borevitz, Norman Warthmann.

**Writing – review & editing:** Kevin D. Murray, Christfried Webers, Cheng Soon Ong, Justin Borevitz, Norman Warthmann.

## References

1. Metzker ML. Sequencing Technologies—the next Generation. Nature Reviews Genetics. 2010; 11 (1):31–46. https://doi.org/10.1038/nrg2626 PMID: 19997069
2. The Arabidopsis Genome Iniative. Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana. Nature. 2000; 408(6814):796–815. https://doi.org/10.1038/35048692 PMID: 11130711
3. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, et al. The Norway Spruce Genome Sequence and Conifer Genome Evolution. Nature. 2013; 497(7451):579–584. https://doi.org/10.1038/nature12211 PMID: 23698360
4. Brandt DYC, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Project Phase I Data. G3: Genes|Genomes|Genetics. 2015; 5(5):931–941. https://doi.org/10.1534/g3.114.015784 PMID: 25787242
5. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De Novo Assembly and Genotyping of Variants Using Colored de Bruijn Graphs. Nature Genetics. 2012; 44(2):226–232. https://doi.org/10.1038/ng.1028 PMID: 22231483
6. Bergelson J, Buckler ES, Ecker JR, Nordborg M, Weigel D. A Proposal Regarding Best Practices for Validating the Identity of Genetic Stocks and the Effects of Genetic Variants. The Plant Cell. 2016; 28(3):606–609. https://doi.org/10.1105/tpc.15.00502 PMID: 26956491
7. Harlan JR, de Wet JMJ. The Compilospecies Concept. Evolution. 1963; 17(4):497. https://doi.org/10.1111/j.1558-5646.1963.tb03307.x
8. Song K, Ren J, Reinert G, Deng M, Waterman MS, Sun F. New Developments of Alignment-Free Sequence Comparison: Measures, Statistics and next-Generation Sequencing. Briefings in Bioinformatics. 2014; 15(3):343–353. https://doi.org/10.1093/bib/bbt067 PMID: 24064230
9. Tang J, Hua K, Chen M, Zhang R, Xie X. A Novel k-Word Relative Measure for Sequence Comparison. Computational Biology and Chemistry. 2014; 53, Part B:331–338. https://doi.org/10.1016/j.compbiolchem.2014.10.007
10. Forêt S, Wilson SR, Burden CJ. Characterizing the D2 Statistic: Word Matches in Biological Sequences. Statistical Applications in Genetics and Molecular Biology. 2009; 8(1):1–21. https://doi.org/10.2202/1544-6115.1447
11. Sims GE, Jun SR, Wu GA, Kim SH. Alignment-Free Genome Comparison with Feature Frequency Profiles (FFP) and Optimal Resolutions. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(8):2677–2682. https://doi.org/10.1073/pnas.0813249106 PMID: 19188606
12. Leimeister CA, Morgenstern B. Kmacs: The k-Mismatch Average Common Substring Approach to Alignment-Free Sequence Comparison. Bioinformatics. 2014; p. btu331. https://doi.org/10.1093/bioinformatics/btu331
13. Leimeister CA, Boden M, Horwege S, Lindner S, Morgenstern B. Fast Alignment-Free Sequence Comparison Using Spaced-Word Frequencies. Bioinformatics. 2014; p. btu177. https://doi.org/10.1093/bioinformatics/btu177
14. Yi H, Jin L. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. Nucleic acids research. 2013; 41(7):e75–e75. https://doi.org/10.1093/nar/gkt003 PMID: 23335788
15. Vinga S. Information Theory Applications for Biological Sequence Analysis. Briefings in Bioinformatics. 2014; 15(3):376–389. https://doi.org/10.1093/bib/bbt068 PMID: 24058049

16. Morgenstern B, Zhu B, Horwege S, Leimeister CA. Estimating Evolutionary Distances between Genomic Sequences from Spaced-Word Matches. Algorithms for Molecular Biology. 2015; 10(1):5. https://doi.org/10.1186/s13015-015-0032-x PMID: 25685176

17. Aflitos SA, Severing E, Sanchez-Perez G, Peters S, de Jong H, de Ridder D. Cnidaria: Fast, Reference-Free Clustering of Raw and Assembled Genome and Transcriptome NGS Data. BMC Bioinformatics. 2015; 16:352. https://doi.org/10.1186/s12859-015-0806-7 PMID: 26525298

18. Fan H, Ives AR, Surget-Groba Y, Cannon CH. An Assembly and Alignment-Free Method of Phylogeny Reconstruction from next-Generation Sequencing Data. BMC Genomics. 2015; 16(1):522. https://doi.org/10.1186/s12864-015-1647-5 PMID: 26169061

19. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: Fast Genome and Metagenome Distance Estimation Using MinHash. Genome Biology. 2016; 17:132. https://doi.org/10.1186/s13059-016-0997-x PMID: 27323842

20. Reinert G, Chew D, Sun F, Waterman MS. Alignment-Free Sequence Comparison (I): Statistics and Power. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology. 2009; 16(12):1615–1634. https://doi.org/10.1089/cmb.2009.0198

21. Wan L, Reinert G, Sun F, Waterman MS. Alignment-Free Sequence Comparison (II): Theoretical Power of Comparison Statistics. Journal of Computational Biology. 2010; 17(11):1467–1490. https://doi.org/10.1089/cmb.2010.0056 PMID: 20973742

22. Allman ES, Rhodes JA, Sullivant S. Statistically-Consistent k-Mer Methods for Phylogenetic Tree Reconstruction. arXiv:151101956 [q-bio]. 2015;

23. Ren J, Song K, Deng M, Reinert G, Cannon CH, Sun F. Inference of Markovian Properties of Molecular Sequences from NGS Data and Applications to Comparative Genomics. Bioinformatics. 2016; 32(7):993–1000. https://doi.org/10.1093/bioinformatics/btv395 PMID: 26130573

24. Song K, Ren J, Zhai Z, Liu X, Deng M, Sun F. Alignment-Free Sequence Comparison Based on Next-Generation Sequencing Reads. Journal of Computational Biology. 2013; 20(2):64–79. https://doi.org/10.1089/cmb.2012.0228 PMID: 23383994

25. Jiang B, Song K, Ren J, Deng M, Sun F, Zhang X. Comparison of Metagenomic Samples Using Sequence Signatures. BMC Genomics. 2012; 13:730. https://doi.org/10.1186/1471-2164-13-730 PMID: 23268604

26. Cormode G, Muthukrishnan S. An improved data stream summary: the count-min sketch and its applications. Journal of Algorithms. 2004; 55(1):58–75. https://doi.org/10.1016/j.jalgor.2003.12.001

27. Zhang Q, Pell J, Canino-Koning R, Howe AC, Brown CT. These Are Not the K-Mers You Are Looking For: Efficient Online K-Mer Counting Using a Probabilistic Data Structure. PLoS ONE. 2014; 9(7):e101271. https://doi.org/10.1371/journal.pone.0101271 PMID: 25062443

28. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, et al. The Khmer Software Package: Enabling Efficient Nucleotide Sequence Analysis. F1000Research. 2015 https://doi.org/10.12688/f1000research.6924.1 PMID: 26535114

29. Sims GE, Jun SR, Wu GA, Kim SH. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. Proceedings of the National Academy of Sciences. 2009; 106(40):17077–17082. https://doi.org/10.1073/pnas.0909377106

30. Hofmann T, Schölkopf B, Smola AJ. Kernel Methods in Machine Learning. The Annals of Statistics. 2008; 36(3):1171–1220. https://doi.org/10.1214/009053607000000677

31. Guennebaud G, Jacob B, others. Eigen V3; 2010.

32. Li JY, Wang J, Zeigler RS. The 3,000 Rice Genomes Project: New Opportunities and Challenges for Future Rice Research. GigaScience. 2014; 3(1):8. https://doi.org/10.1186/2047-217X-3-8 PMID: 24872878

33. The 3,000 rice genomes project. The 3,000 Rice Genomes Project. GigaScience. 2014; 3(1):7. https://doi.org/10.1186/2047-217X-3-7 PMID: 24872877

34. Flowers JM, Hazzouri KM, Pham GM, Rosas U, Bahmani T, Khraiwesh B, et al. Whole-Genome Resequencing Reveals Extensive Natural Variation in the Model Green Alga Chlamydomonas Reinhardtii. The Plant Cell. 2015; 27(9):2353–2369. https://doi.org/10.1105/tpc.15.00492 PMID: 26392080

35. Edwards J, Johnson C, Santos-Medellín C, Lurie E, Podishetty NK, Bhatnagar S, et al. Structure, Variation, and Assembly of the Root-Associated Microbiomes of Rice. Proceedings of the National Academy of Sciences. 2015; 112(8):E911–E920. https://doi.org/10.1073/pnas.1414592112

36. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. arXiv:12034802 [q-bio]. 2012

37. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A High-Performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. Bioinformatics. 2012; 28(24):3326–3328. https://doi.org/10.1093/bioinformatics/bts606 PMID: 23060615

38. McVean G. A Genealogical Interpretation of Principal Components Analysis. PLOS Genet. 2009; 5(10): e1000686. https://doi.org/10.1371/journal.pgen.1000686 PMID: 19834557

39. Melsted P, Pritchard JK. Efficient Counting of k-Mers in DNA Sequences Using a Bloom Filter. BMC bioinformatics. 2011; 12:333. https://doi.org/10.1186/1471-2105-12-333 PMID: 21831268

40. Salton G, Buckley C. Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management. 1988; 24(5):513–523. https://doi.org/10.1016/0306-4573(88)90021-0

41. Chen S, Goodman J. An Empirical Study of Smoothing Techniques for Language Modeling. In: Proceedings of the 34th Annual Meeting of the ACL; 1996. p. 310–318.

42. Leslie C, Kuang R. Fast String Kernels Using Inexact Matching for Protein Sequences. J Mach Learn Res. 2004; 5:1435–1455.

43. Leslie CS, Eskin E, Cohen A, Weston J, Noble WS. Mismatch String Kernels for Discriminative Protein Classification. Bioinformatics (Oxford, England). 2004; 20(4):467–476. https://doi.org/10.1093/bioinformatics/btg431

44. Leslie C, Eskin E, Noble WS. The Spectrum Kernel: A String Kernel for SVM Protein Classification. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2002; p. 564–575.

45. Gog S, Beller T, Moffat A, Petri M. From Theory to Practice: Plug and Play with Succinct Data Structures. In: Gudmundsson J, Katajainen J, editors. Experimental Algorithms: 13th International Symposium, SEA 2014, Copenhagen, Denmark, June 29—July 1, 2014. Proceedings. Cham: Springer International Publishing; 2014. p. 326–337.

46. Solomon B, Kingsford C. Fast Search of Thousands of Short-Read Sequencing Experiments. Nature Biotechnology. 2016; 34(3):300–302. https://doi.org/10.1038/nbt.3442 PMID: 26854477

47. Brachi B, Morris GP, Borevitz JO. Genome-Wide Association Studies in Plants: The Missing Heritability Is in the Field. Genome biology. 2011; 12(10):232. https://doi.org/10.1186/gb-2011-12-10-232 PMID: 22035733

48. Spindel JE, McCouch SR. When More Is Better: How Data Sharing Would Accelerate Genomic Selection of Crop Plants. New Phytologist. 2016; p. n/a–n/a

49. Köster J, Rahmann S. Snakemake—a Scalable Bioinformatics Workflow Engine. Bioinformatics. 2012; 28(19):2520–2522. https://doi.org/10.1093/bioinformatics/bts480 PMID: 22908215

50. Perez F, Granger BE. IPython: A System for Interactive Scientific Computing. Computing in Science & Engineering. 2007; 9(3):21–29. https://doi.org/10.1109/MCSE.2007.53

51. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks—a Publishing Format for Reproducible Computational Workflows. In: Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing. IOS Press; 2016. p. 87.

52. Staab PR, Zhu S, Metzler D, Lunter G. Scrm: Efficiently Simulating Long Sequences Using the Approximated Coalescent with Recombination. Bioinformatics. 2015; 31(10):1680–1682. https://doi.org/10.1093/bioinformatics/btu861 PMID: 25596205

53. Cartwright RA. DNA Assembly with Gaps (Dawg): Simulating Sequence Evolution. Bioinformatics. 2005; 21(Suppl 3):iii31–iii38. https://doi.org/10.1093/bioinformatics/bti1200 PMID: 16306390

54. Holtgrewe M. Mason—A Read Simulator for Second Generation Sequencing Data. Technical Report FU Berlin. 2010;.

55. Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python; 2001–.

56. Murray K. SRApy: Pythonic Tools for Accessing the Short Read Archive. Zenodo. 2016;

57. Joshi NA, Fass JN. Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files; 2011.

58. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY: Springer New York; 2009.

59. R Core Team. R: A Language and Environment for Statistical Computing; 2016. Available from: https://www.R-project.org/.

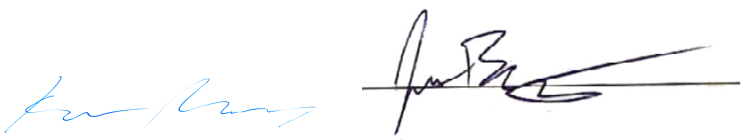
60. Li H. Seqtk—Toolkit for Processing Sequences in FASTA/Q Formats; 2008. https://github.com/lh3/seqtk.

# Chapter 5

# Global Diversity of the *Brachypodium* Species Complex as a Resource for Genome-Wide Association Studies Demonstrated for Agronomic Traits in Response to Climate

This chapter describes a large, collaborative project to establish the genetic resources required for genome-wide association studies (GWAS) in *Brachypodium*. I contributed to many analyses and experiments presented in this article, namely: I designed and conducted the Genotyping-by-sequencing analysis, I performed several population genetic analyses (e.g. accession clustering, LD calculation), I conducted the whole-genome genotyping experiment, co-designed the environmental growth conditions and control software, and I assisted with the result interpretation and writing the manuscript.

This work has been published in Genetics (2019; doi: 10.1534/genetics.118.301589). The senior author authorises the inclusion of this manuscript in my thesis.



# Global Diversity of the Brachypodium Species Complex as a Resource for Genome-Wide Association Studies Demonstrated for Agronomic Traits in Response to Climate

**Pip B. Wilson,*[,1,2] Jared C. Streich,*[,2,3] Kevin D. Murray,*[,2] Steve R. Eichten,* Riyan Cheng,*[,†] Nicola C. Aitken,*[,‡] Kurt Spokas,[§] Norman Warthmann,* Sean P. Gordon,** Accession Contributors,[4] John P. Vogel,** and Justin O. Borevitz*[,5]**

*The ARC Centre of Excellence in Plant Energy Biology and [‡]Ecogenomics and Bioinformatics Lab, Research School of Biology, Australian National University, Canberra, Australian Capital Territory 200, Australia, [†]Department of Psychiatry, University of California San Diego, La Jolla, California 92093, [§]Soil and Water Management, Agricultural Research Service, United States Department of Agricutlture (USDA), St. Paul, Minnesota 55108, and **Department of Energy, Joint Genome Institute, Walnut Creek, California 94598

ORCID IDs: 0000-0003-4861-1188 (J.C.S.); 0000-0001-8408-3699 (J.O.B.)

**ABSTRACT** The development of model systems requires a detailed assessment of standing genetic variation across natural populations. The Brachypodium species complex has been promoted as a plant model for grass genomics with translation to small grain and biomass crops. To capture the genetic diversity within this species complex, thousands of Brachypodium accessions from around the globe were collected and genotyped by sequencing. Overall, 1897 samples were classified into two diploid or allopolyploid species, and then further grouped into distinct inbred genotypes. A core set of diverse *B. distachyon* diploid lines was selected for whole genome sequencing and high resolution phenotyping. Genome-wide association studies across simulated seasonal environments was used to identify candidate genes and pathways tied to key life history and agronomic traits under current and future climatic conditions. A total of 8, 22, and 47 QTL were identified for flowering time, early vigor, and energy traits, respectively. The results highlight the genomic structure of the Brachypodium species complex, and the diploid lines provided a resource that allows complex trait dissection within this grass model species.

**KEYWORDS** population genetics; climate change; agronomic traits; climate simulation; genome-wide association studies; ecogenomics; Brachypodium distachyon; genotyping; plant physiology


doi: https://doi.org/10.1534/genetics.118.301589
Manuscript received September 9, 2018; accepted for publication November 8, 2018; published Early Online November 16, 2018.
Available freely online through the author-supported open access option.

Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7345160.
[1]Present address: Grains Research and Development Corporation, Canberra, ACT, Australia.
[2]These authors contributed equally to this work.
[3]Present address: Oak Ridge National Lab, Oak Ridge, TN 37830.
[4]List of Accession Contributors in the Acknowledgments.
[5]Corresponding author: Australian National University, Bldg. 134, Linnaeus Way, Canberra, ACT 200, Australia. E-mail: Justin.borevitz@anu.edu.au

CLIMATE change is impacting the production of food worldwide (Wheeler and von Braun 2013), and increasing global demand will soon outstrip the rate of improvement in crop yield by traditional breeding methods (Ray *et al.* 2013). To address food and climate security, there is a need for agricultural innovation across a range of scientific disciplines, from genomics to phenomics in new species across the landscape (Rivers *et al.* 2015). Breeding for more variable future climates, and for broad adaptability, requires an understanding of the plasticity of the genetic architecture of agronomic traits across environments. The use of dynamic climate chambers, that can mimic regional diurnal and seasonal climate types (Brown *et al.* 2014), allows us to examine the genetic architecture underlying complex adaptive traits across field-like environments.

Three complex traits that have a large impact on yield are ear emergence, early vigor, and energy use efficiency. The timing of ear emergence is crucially important to yield in many grain-growing regions, including Australia, where early flowering may lead to cold-induced sterility, while late flowering may result in heat stress or lack of water-limiting grain filling. Early vigor, defined as an increase in the above-ground biomass prior to stem elongation, is a beneficial trait in many environment types, especially when combined with increased transpiration efficiency (Condon *et al.* 2004). Since vapor pressure is low in winter, increased biomass during early growth improves plant water use efficiency. Early vigor also increases competition against weeds, reduces soil evaporation and may improve yields by increasing total seasonal biomass (Wilson *et al.* 2015a). Energy use efficiency is a relatively understudied component of plant growth that represents the efficient transfer of energy, acquired through photosynthesis, to the grain, and may significantly affect yield. Early studies indicate that energy efficiency, via lower respiration rates, is correlated with an increase in biomass in monocot species (Wilson and Jones 1982; Winzeler *et al.* 1988). Identification of the genetic architecture of energy use efficiency, timing of heading, and early vigor traits, as well as the genetic sensitivity to future temperature profiles, could accelerate breeding in crop species via selection for improved predicted yields in the future.

Genome-wide association studies (GWAS) combine dense genetic markers, identified via next-generation sequencing and high-throughput phenotyping, to identify the causative alleles and to predict complex quantitative traits (Atwell *et al.* 2010). The improvement of crop yield involves many complex traits, and the expression of these traits can be highly dependent on the growth environment. GWAS is an excellent method for mapping and predicting yield-related traits and their interaction with the environment. GWAS has been undertaken in a number of crop species; for dozens of agronomic traits in diploid species such as rice, barley, and corn (for review, see Huang and Han 2014), and has even be used reasonably successfully in wheat despite the added complexity of a hexaploid genome (*e.g.*, Sukumaran *et al.* 2014).

*Brachypodium distachyon* is a model species for temperate C3 grass crops such as wheat, barley, rye, and oats as it is also located in the Pooideae family and has a number of advantageous characteristics as a model species (Draper *et al.* 2001; Garvin *et al.* 2008; Mur *et al.* 2011; Brutnell *et al.* 2015). *B. distachyon* also has a number of advantages over the related domestic Pooideae for a GWAS approach as it is a wild species with a wide climatic distribution, resulting in diverse phenotypes, as well as wide genomic diversity, for traits involved in life strategy and abiotic stress tolerance. *B. distachyon* has a small, fully sequenced genome of 270 Mb (The International Brachypodium Initiative 2009) compared to the 16 Gb of wheat (The International Wheat Genome Sequencing Consortium 2017) or 5.1 Gb of barley (The International Brachypodium Initiative 2009). It also contains a low percentage of repetitive noncoding DNA at 21.4% of nucleotides compared to >80% in wheat (Wicker *et al.* 2011) and 84% in barley (The International Barley Genome Sequencing Consortium 2012). This means that sequence reads from *B. distachyon* are much easier to identify and align compared to wheat, with a larger proportion of the sequencing providing useful reads. Finally, and perhaps most importantly, the short stature of *B. distachyon* allows large numbers of plants to be taken through full life cycles in controlled growth conditions.

*Brachypodium* is widespread throughout temperate regions, including its native Mediterranean range and introduced range in Australia, South Africa, and the western United States (Vogel *et al.* 2009; Wilson and Jones 2015). A large number of accessions have been collected throughout the world by the *Brachypodium* community, but the use of these collections in genomic association studies has been delayed by the cryptic nature of the *Brachypodium* species complex. The three species in this complex are difficult to distinguish in the field and include the diploid *B. distachyon*, the diploid *B. stacei*, and the allotetraploid *B. hybridum*, which contains one *B. distachyon*-like genome and one *B. stacei*-like genome (Hasterok *et al.* 2004; Catalán *et al.* 2012; Idziak *et al.* 2014). To add to the complexity, there is evidence of distinct subgroups of *B. distachyon* (Hasterok *et al.* 2004; Catalán *et al.* 2012; Idziak *et al.* 2014; Tyler *et al.* 2016). While the genome of the Bd21 reference genotype of *B. distachyon* was published in 2010, the genome of *B. stacei* and other SNP corrected genomes were released online in 2016 (DOE-JGI, http://phytozome.jgi.doe.gov/). Recently, a *B. distachyon* pan genome was published identifying geographic diversity and many new genes not identified in the initial reference (Gordon *et al.* 2017). Prior to our study, species identification has commonly been undertaken by morphoanatomical classification, a small number of markers, or cytology (*e.g.*, Hasterok *et al.* 2004; The International Wheat Genome Sequencing Consortium 2017). There is a need for a rapid identification of species, subgroup, and genotype lineages within the *Brachypodium* species complex to aid the selection of HapMap sets, and to enable landscape genomic studies of migration and adaptation.

In this study, we aimed to (1) characterize the species, genotype, and population structure of a *Brachypodium* global diversity set to select a core haplotype mapping set for GWAS in *B. distachyon* and (2) identify the genetic architecture and plasticity of the agriculturally relevant traits of heading date, early vigor, and energy use efficiency in response to climate.

## Materials and Methods

### *Genotyping by sequencing and species identification*

Genotyping by sequencing (GBS) was undertaken as described by Elshire and colleagues (Elshire *et al.* 2011) using *Pst*I enzyme and a library of homemade barcoded adaptors (see https://github.com/borevitzlab/brachy-genotyping; Morris *et al.* 2011; Nicotra *et al.* 2016). Approximately 384 samples were multiplexed to run on a single lane in an Illumina HiSeq 2000 with a median number of 564,000 100-bp read pairs

per sample (https://github.com/borevitzlab/brachy-genotyping). Sequencing runs were undertaken by the Biomolecular Resource Facility [The John Curton School of Medical Research (JCSMR), Australian National University (ANU)].

Axe (Murray and Borevitz 2018) was used to demultiplex sequencing lanes into libraries, allowing no mismatches. AdapterRemoval (Schubert *et al.* 2016) was used to remove contaminants from reads, and merge overlapping read pairs. Reads were aligned using BWA MEM (Li 2013; Li and Durbin 2009) to the Bd21-3 (*B. distachyon*) and ABR114 (*B. stacei*) reference genomes (Phytozome v.12.1), and to a *B. hybridum* pseudoreference genome created by concatenating the *B. stacei* and *B. distachyon* reference genomes (Supplemental Material, File S1). Variants were called using the multiallelic model of samtools mpileup (Li 2011) and bcftools call (Danecek *et al*. 2016). Variants were filtered with bcftools filter, keeping only SNPs of reasonable mapping and variant qualities ($\geq$10) and sequencing depth across samples ($\geq$5 reads across all samples).

To determine the species of each of the accessions, we computed the proportion of each chromosome in the *B. hybridum* pseudoreference covered with at least three reads, excluding reads that mapped to multiple locations in the pseudoreference, using mosdepth (Pedersen and Quinlan 2017). The proportions of the *B. distachyon/B. stacei* genomes covered were normalized to be in [0, 1], and then used to assign samples into threshold groups: *B. stacei* ($<$0.03), intermediate *B. stacei/B. hybridum* ($<$0.28), *B. hybridium* ($<$0.34), intermediate *B. hybridum/B. distachyon* (0.94), and *B. distachyon* ($>$0.94); an additional group consisted of low coverage samples ($<$100,000 reads in total). Samples from intermediate and low coverage groups were excluded, and only variants in the respective genomes were used to allocate the three species groups.

### Population structure of B. distachyon

To determine the population structure of *B. distachyon*, a pairwise identity-by-state (IBS) genetic distance was calculated to identify, among 490 high-quality samples, a core diversity set of 72 distinct genotypes using 82,800 SNPs derived from GBS data and the SNPRelate package using a z-score of 3.5. Occasionally, when genotypes are closely related, noise between technical replicates of an accession will result in them being split across the related genotypes. Therefore, we keep replicate(s) from the genotype with the majority of replicates for that accession, breaking ties by keeping the replicate with the lowest missing data. In addition, 29 accessions whose geographic origin was suspect were also excluded.

To avoid bias from including up to 30 inbred accessions of the same genotype, a reduced set was input into STRUCTURE V.2.3.4 (Evanno *et al.* 2005). A total of six replicates were run of population (K) 1–13 with a burn-in setting of 10,000 sets, and 100,000 permutations per run (Figure 1B and File S3). The optimal K was determined as $K = 3$ by Evanno's Delta K, processed via Structure Harvester and CLUMPP (Evanno *et al*. 2005, Jakobsson and Rosenberg 2007; Earl and vonHoldt 2012). Barplots and pie charts were generated via inhouse developed R scripts available through github (https://github.com/borevitzlab/brachy-genotyping-notes).

For *B. distachyon*, the pairwise distance between genotypes was also calculated in R and plotted as a dendrogram (File S2). From this, a set of 107 accessions were selected to represent the genotypic diversity of the species for whole genome sequencing (WGS) to maximize SNP coverage across the genome.

### Whole genome sequencing

For WGS, sequencing libraries for individual samples were prepared from 6 ng genomic DNA with the Nextera DNA Library Prep kit (Illumina, San Diego, CA). Libraries were enriched and barcoded with custom i5-, and i7-compatible oligos and Q5 High-Fidelity DNA Polymerase (NEB, Ipswich, MA). Libraries were pooled and sequenced in one lane on a NextSeq 500 sequencer (Illumina).

Trimit (Murray and Borevitz 2017) was used to clean WGS reads of adaptors, and merge overlapping read pairs. BWA MEM was then used to align these reads against the Bd21-1 reference genome (version 314_v3.1; The International Brachypodium Initiative 2009). Variants were called using freebayes (Garrison and Marth 2012) with default parameters. Variants were filtered such that only variants meeting the following criteria were kept: variant quality $>$20, minor allele frequency $\geq$2%. Heterozygous variant calls were changed to missing; due to the inbred nature of these accessions, heterozygous calls were almost certainly erroneous (https://github.com/borevitzlab/brachy-genotyping).

Linkage disequilibrium (LD) was calculated across the *B. distachyon* genome using consecutive windows of 2000 SNPs from the whole genome data of the HapMap 74 set (http://github.com/borevitzlab/brachy-genotyping-notes).

### Plant growth

Individual grain of each genotype was planted 2.5 cm deep in square plastic pots (5 cm width, 8 cm deep) in a mix of 50:50 soil:washed river sand that had been steam pasteurized. Pots were then placed at 4° in the dark for 3 days to stratify the seed before being moved to specially modified climate chambers (see Garrison and Marth 2012). Accessions were organized in a randomized block design, in trays of 20 plants. The chambers have been fitted with seven LED light panels and are controlled to change the light intensity, light spectrum, air temperature and humidity every 5 min. Seasonal changes in climatic conditions and photoperiod were modeled using SolarCalc software (Spokas and Forcella 2006). The Wagga Wagga region is centered on $\sim -35$S, 147E with an elevation of 147 m. Plants were fertilized with Thrive (N:P:K 25:5:8.8 + trace elements, Yates) and watered with tap water as needed. Growth stages were recorded based on the Huan developmental stage (Haun *et al*. 1973) up until stem elongation and thereafter the Zadoks scale was

used. Total leaf area was measured with a Li-1300 Area Meter (Li-COR). For dry weight, leaf tissue was dried in a paper envelope at 60° for 5 days before weighing.

### Conversions of phenotypic data

Thermal time was calculated from the logged condition within each chamber with the following formula:

$$\text{If Temp.\_1} > 2°\text{C, then TT\_2} = \text{TT\_1} + [(\text{Temp.\_2} - 2) \times \Delta\text{Time}]\text{\_}(2 - 1)$$

where $TT_i$ is accumulated thermal time at a particular timepoint $i$, and $\text{Temp.}_i$ is the air temperature at a particular timepoint $i$.

Photothermal units (PTU) were calculated using the logged data from a photosynthetically active radiation (PAR) sensor in the middle of the chamber and the following formula:

$$(\text{PTU})\text{\_}1 = (\text{TT})\text{\_}1 \times (\text{PAR})\text{\_}1$$

where TT is the accumulated thermal time at timepoint $i$ and PAR is the measured photosynthetically active radiation at timepoint $i$.

Growth rates (GR) were calculated as:

$$\text{GR} = [(\Delta\text{GS})\text{\_}(T\text{\_}2(-T)\text{\_}1)]/[\Delta(\text{Time})\text{\_}((T\text{\_}2 - T)\text{\_}1)]$$

where GS is the Huan growth stage and $T_1$ was about one leaf for the initial linear growth stage (GR1), $T_1$ was about one leaf and three leaves, and the faster growth stage (GR2) between three leaves and five leaves. The Phyllachron interval, the time taken to grow one leaf, was calculated as:

$$\text{Phyllachron Interval} = (T\text{\_}2 - T)\text{\_}e/(\text{GS})\text{\_}2$$

where $T_2$ is the unit of time at about the three-leaf stage and $T_e$ is the unit of time at seedling emergence for that particular plant. $GS_2$ is the Huan growth stage at $T_2$.

Final growth efficiency was calculated when plants reached ear emergence. The final growth efficiency 1 was calculated as:

$$\text{Final growth efficiency 1} = (\text{Biomass at ear emergence } (g))/(\Delta\text{Thermal time})$$

where accumulated thermal time is calculated from seedling emergence to ear emergence. The final growth efficiency 2 was calculated as:

$$\text{Final growth efficiency 2} = (\text{Biomass at ear emergence } (g))/(\Delta\text{Photothermal units})$$

where accumulated photothermal units is calculated from seedling emergence to ear emergence.

### Energy use efficiency traits

Energy use efficiency traits were measured on plants from the 2015–2050 Temperature experiment at a four- to five-leaf stage. Photosynthetic parameters were measured using a Trayscan system (PSI) incorporating pulse amplitude modification (PAM) chlorophyll fluorescence measures of quantum efficiency (Rungrat *et al.* 2016). The parameters measured included photosynthetic efficiency, nonphotochemical quenching and photo-inhibition. See File S2 for protocol.

Dark respiration rate was measured using the Q2 system (Astec Global) as in Scafaro *et al.* (2017). In brief, this system uses an oxygen-sensitive fluorescent dye embedded in a cap to monitor the oxygen depletion with a tube containing the sample. A 3 cm fragment in the center of the last fully expanded leaf of each plant was used to measure dark respiration per unit area and per unit dry mass.

Several energy use efficiency formulas were calculated. These included a ratio of dark respiration to photosynthesis, and measures of growth per unit dark respiration. These were as follows:

$$\text{Energy use efficiency 1} = 1 - (\text{respiration per unit area})/(\text{average photosynthetic efficiency})$$

$$\text{Energy use efficiency 2} = (\text{seedling height})/(\text{respiration per g dry weight})$$

$$\text{Energy use efficiency 3} = (\text{leaf \#3 length})/(\text{respiration per g dry weight})$$

$$\text{Energy use efficiency 4} = (\text{seedling height})/(\text{respiration per unit area})$$

### Heritability

Broad-sense heritability was calculated from the phenotype data using the nlme package in R.

### GWAS analysis

In preparation for GWAS, the genotype data were filtered to remove nonvariant SNPs and redundant SNPs (*i.e.*, SNPs whose genotypes are not different from adjacent SNPs but have more missing data points). Then, SNPs with a minor allele frequency of <3% were filtered out. As there was 18.5% missing data in the original data set, imputation was undertaken. First, if the observed genotypes of two adjacent SNPs were not different, then the missing genotype of one SNP was replaced by the observed genotype of the other SNP. Second, the nearest neighbor (NN) method was implemented to impute the remaining missing genotypes based on Huang *et al.* (2010) with some modifications. The nearest 50 SNPs from each side of the SNP under imputation were selected to estimate similarity between each pair of accessions, and then the missing genotype of an accession was replaced by the observed majority genotype of the closest five accessions.

These parameters were determined by simulations to achieve an optimal imputation success rate, which was 97.95% for our data. Finally, SNPs with a minor allele frequency <5% were filtered. For the phenotype data, the average value for the four replicates of each accession was calculated.

Linear mixed-effect models were employed to identify genetic variants underlying phenotypes of interest

$$y = x\beta + z\gamma + u + \epsilon$$

where $\boldsymbol{y} = (y_1, y_2, \ldots\ldots, y_n)'$ denotes phenotypic values, $\boldsymbol{x} = (x_{ij})_{nx(k+1)}$ represents intercept and $k$ covariates (if any) with effects $\boldsymbol{\beta}$, $\boldsymbol{z}$ is a vector of the coded genotypes at a scanning locus with effect $\gamma$, $\boldsymbol{u} = (u_1, u_2, \ldots\ldots, u_n)'$ represents polygenic variation, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots\ldots, \epsilon_n)$ the residual effect. It was assumed that $\boldsymbol{u} \sim N(0, \boldsymbol{K}\sigma_g^2)$, $\boldsymbol{\epsilon} \sim N(0, I^2)$ and $\boldsymbol{u}$ was independent of $\boldsymbol{\epsilon}$. The genetic relationship matrix $\boldsymbol{K}$ was estimated by IBS from genotypic data with markers on the chromosome under scan being excluded to avoid proximal contamination (Listgarten *et al*. 2012; Cheng *et al*. 2013). Estimation of $\boldsymbol{K}$ and genome scan were performed in R package QTLRel (Cheng *et al*. 2011).

To determine a significance threshold, the permutation test was implemented on 1000 permutations of the phenotype data to estimate the genome-wide significance threshold at 0.05 for the trait of days to ear emergence. The significance threshold was determined to be a LOD (logarithm of odds) of 4.43583.

### *Data availability*

GBS and whole genome sequence data are available in the sequence read archive at NCBI, BioprojectID PRJNA505390. Supplemental Figures and tables are available in FigShare. Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7345160.

## Results

### *Cryptic Brachypodium species, diverse genotypes, and population structure identified using GBS*

To establish a diverse set of germplasm, thousands of *Brachypodium* accessions were collected on trips to southwest Europe, south-eastern Australia, the western USA, and through collaborations with the international *Brachypodium* community (https://github.com/borevitzlab/brachy-genotyping/blob/master/metadata/brachy-metadata.csv). Out of these, 1968 accessions were grown to produce single-seed descent lines in the greenhouses at the ANU for subsequent genomic analysis. A reduced representation approach, *PstI* digest, GBS was used to genetically profile the accessions.

Although once described as a single species, *B. distachyon* has more recently been shown to exist as a species complex consisting of a 5 chromosome *B. distachyon*, 10 chromosome *B. stacei*, and a 15 chromosome allopolyploid *B. hybridum* (Catalán *et al*. 2012). To categorize each accession into species within the *Brachypodium* complex, GBS tags were mapped to a merged reference genome consisting of *B. distachyon* (Bd21-3) and *B. stacei* (ABR114) (v1.1 DOE-JG, https://phytozome.jgi.doe.gov). Most accessions were readily distinguished as having reads that aligned to either or both reference genomes (see *Materials and Methods*; File S1). The majority of accessions, 56% (1100/1968), were identified as *B. hybridum*. In contrast, only 3% (60/1968) were classified as *B. stacei*, while 35% (698/1968) were *B. distachyon*. The remaining 6% (110/1968) could not be definitively assigned. Mapping of the accessions' geographic locations showed that *B. hybridum* has expanded across the globe, representing essentially all the collections outside the native range (Figure 1A). Conversely, *B. distachyon* is largely limited to the native Mediterranean and Western Asian regions, with *B. stacei* in the same area, but less common.

Due to the highly selfing nature of all *Brachypodium* species, we next sought to categorize accessions into unique whole genome genotypes representing a single inbred lineage. Of the 698 accessions identified as *B. distachyon*, 490 could be reliably genotyped at 81,400 SNPs. We used the SNPRelate package (Zheng *et al.* 2012) to cluster these 490 accessions into 72 genotypes (see *Materials and Methods*; https://github.com/borevitzlab/brachy-genotyping-notes; File S2). Recombinant inbred lines, included as positive controls, were often called as unique genotypes as expected, but were excluded from subsequent analysis of natural population structure.

### *Whole genome variation*

One or two accessions of each unique genotype was selected for further analysis. Whole genome sequencing was performed on this set of 107 *B. distachyon* accessions to determine high density variation at multiple levels, patterns of LD, and to enable GWAS. We identified 2,648,921 SNPs present in at least two accessions. Due to the high inbreeding and clonal family structure observed (File S4), we sought to select a representative accession from each inbred family, reducing 107 accessions to 63 highly diverse genotypes.

Previous genetic analysis on smaller data sets had shown *B. distachyon* to have substantial population structure, forming three groups representing ancestral structure in the Mediterranean region (Filiz *et al.* 2009; Vogel *et al.* 2009; Tyler *et al.* 2016; Gordon *et al.* 2017; Marques *et al.* 2017). To reduce data complexity, SNPs were subsampled to every 100th site to create a final SNP matrix of 26,490 variants that were fed into STRUCTURE v.2.3.4 (Pritchard *et al.* 2000; File S3). STRUCTURE analysis identified three main subgroups among *B. distachyon* genotypes and seven admixed lines (Figure 1B). The yellow lineage was the most diverged and represents subgroup B, with the brown and red structure groups representing the two populations of the A subgroup, split predominantly as an East and West population. Our STRUCTURE clustering is largely consistent with previous results on a smaller, partially overlapping sets of accessions (see Figure 4 of Tyler *et al*. 2016; Gordon *et al*. 2017). To visualize the geographic distribution, the ancestral group composition was summed across accessions for each geographic site (Figure

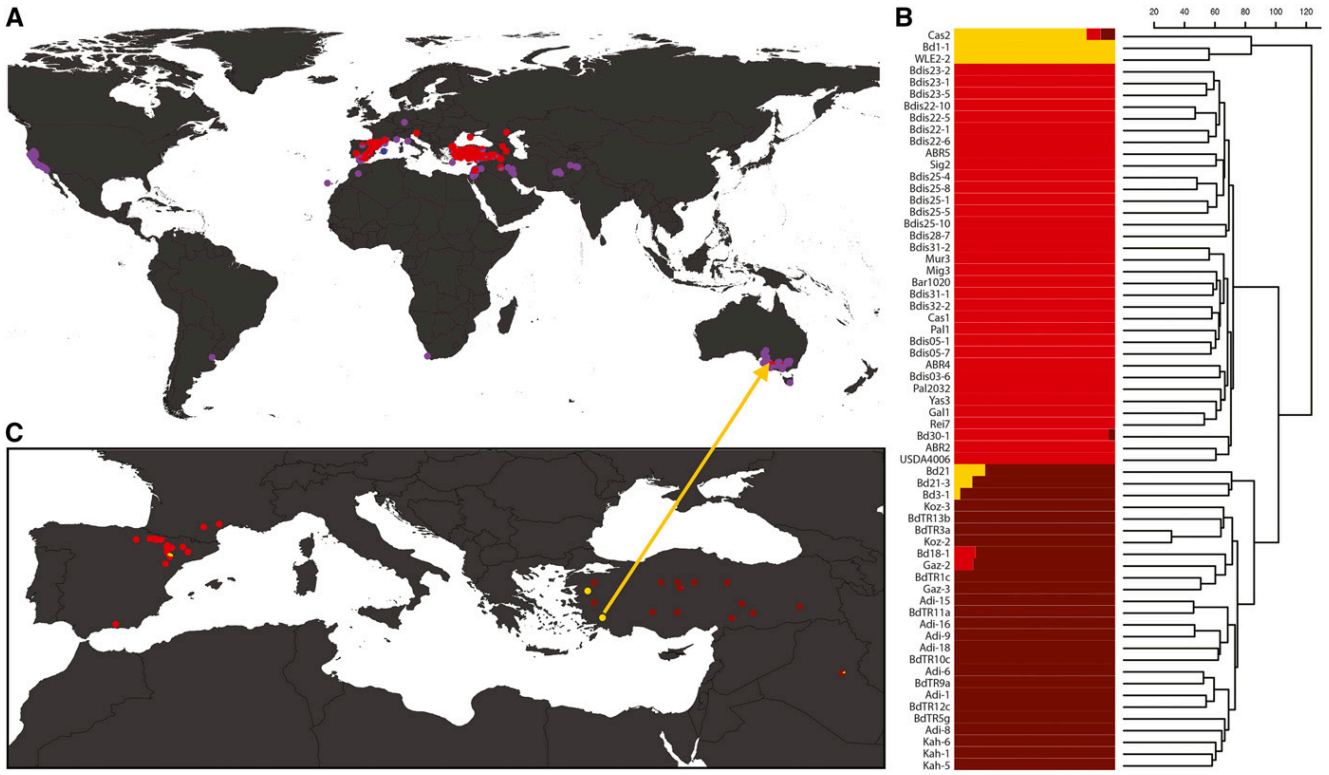



**Figure 1** Distribution and genomic diversity of the *B. distachyon* complex. (A) Geographic distribution of 1858 Brachypodium complex accessions, classified by species: pink = *B. distacyhon*, blue = *B. stacei*, and purple = *B. hybridum*. (B) Population structure of the 63 diverse *B. distachyon* genotypes, $K = 3$. The three structure groups correspond to the B subgroup of *B. distachyon* (yellow), and the eastern (brown) and western (red) Mediterranean populations of the A subgroup of *B. distachyon*; and (C) geographic structure of *B. distachyon* across Iberian Peninsula and Turkish region. Proportions of pies represent the number of each *B. distachyon* subgroup (from B) at each site. The arrow from (C) to (A) shows the Australia *B. distachyon* (WLE2-2) and the near-identical accession from Turkey (BdTR9f).

1C). The single *B. distachyon* accession from Australia, WLE2-2, was nearly identical to BdTR9f (GBS data, File S2) from southern western Turkey, from where it may have originated. It is shown in its ancestral location (Figure 1C, arrow).

Although there were only three accessions in the B subgroup, they diverged from the A subgroup with fixed differences at 6.5% of sites. By comparison, fixed divergence between the two clear subpopulations within the A subgroup was 1.5%. Finally, accessions within the same unique genotype diverged at between 0.1 and 0.4% of SNPs. A balanced set of representative accessions across the genotype lineages within just the A subgroup was selected for further genomic and phenomic analysis (File S4).

LD was calculated for consecutive windows of 2000 SNPs across the genome. There was large variation in LD, as the distance of decay to half maximal r2, across the genome (File S5) with the median LD 113 kb (50–235 kb interquartile range) and the maximum >2.4 Mb.

### *Determining the traits and climatic conditions for GWAS in B. distachyon*

For our GWAS study, we wanted to identify high-throughput nondestructive phenotypic measures with high heritability. We also wanted to determine the best environmental conditions to characterize our traits of interest. Hence, two preliminary experiments were undertaken, one for flowering time and one for early vigor.

Flowering time was chosen as an ideal trait for GWAS as it has high heritability in many species including Arabidopsis (Brachi *et al.* 2010) and barley (Maurer *et al.* 2015). Previous studies of *B. distachyon* revealed that the dependence of flowering time on vernalization and photoperiod varies between accessions (Higgins *et al.* 2010; Ream *et al.* 2014; Bettgenhaeuser *et al.* 2017; Woods *et al.* 2017). This study aimed to identify QTL for earliness *per se* in flowering, *i.e.*, those responsive to the accumulation of thermal time. Hence, a preliminary experiment was undertaken to determine if our conditions could meet the vernalization requirements of all *B. distachyon* accessions, and to determine which accessions had strong vernalization requirements in our conditions. To do this, 266 diverse A- and B-subgroup accessions, with five accessions replicated five to six times, were grown in both a simulated Winter sowing, starting June 1, and a Spring sowing, starting September 1, in Wagga Wagga, NSW, Australia (File S6). Ear emergence was monitored as a surrogate measure for flowering time, as flowering occurs largely within the ear in *B. distachyon* so is hard to accurately record (File S7). Out of the 266 accessions, there were 17 accessions that did

not flower in the Spring condition, indicating a strong vernalization requirement (File S8A). All lines flowered in the Winter condition, indicating that night temperatures of 4° were sufficient to meet vernalization requirement. As expected, days to ear emergence showed a strong heritability in the Winter condition, as calculated from the replicated lines ($H^2 = 0.96$). The thermal time to flowering was calculated to determine the dependence of flowering on the accumulation of thermal time. The fast cycling accessions, which did not require vernalization, still required a larger thermal time accumulation than the vernalization requiring accessions (File S8B). This indicates that these either have some low-level requirement for vernalization that is not being fully met in the Spring condition, or that the photoperiod is also a factor in this relationship. As this study aimed to identify QTL for earliness *per se* in flowering, *i.e.*, those responsive to the accumulation of thermal time, we attempted to exclude vernalization and photoperiod effects by focusing on the Winter condition for the GWAS experiment.

In temperate grass crops such as wheat and barley, early vigor can result in an increased yield in short seasons, or in seasons where there is high rainfall (reviewed in Wilson *et al*. 2015c). Often, the dimensions of seedling leaves are measured as a nondestructive surrogate measure for early vigor (Rebetzke and Richards 1999; Wilson *et al.* 2015b). To confirm that this was also an appropriate surrogate measure for early vigor in *B. distachyon*, a highly replicated ($n = 10$) validation experiment was performed on six diverse *B. distachyon* lines (File S9A) in a simulated Wagga Wagga, seasonal climate starting on September 1 (Spring). After 7 weeks, when plants had between four and five mainstem leaves, the dimensions of leaf #3, seedling height, total leaf area, and above-ground dry weight were measured and phenotypic correlations were calculated (File S9B). Broad sense heritability was also calculated to determine which early vigor trait would provide the most power for mapping QTL with GWAS (File S9C). Leaf #3 width and length correlated well with above-ground biomass ($r^2 = 0.46, P < 0.01$, and $r^2 = 0.48, P < 0.01$, respectively) and had quite high heritabilities of $H^2 = 0.60$ and $H^2 = 0.64$, respectively, as compared to above ground dry mass, $H^2 = 0.51$. Interestingly, seedling height also had a strong correlation with above ground biomass ($r^2 = 0.74$, $P < 0.01$) with a heritability of $H^2 = 0.74$. However, this trait was also more highly correlated with developmental stage, as indicated by the number of leaves ($r^2 = 0.21$, $P < 0.01$), than the dimensions of leaf #3. To get the most direct measure of early vigor, without the influence of developmental stage, the dimensions of leaf #3 were chosen as the focus for the GWAS.

### Selection of global HapMap set

High-level population structure confounds GWAS when there are few segregating SNPs in common between ancestral groups relative to variation within each subgroups (Brachi *et al.* 2011). Here, we focused on subgroup A, which contains a majority of unique genotypes, resulting in a HapMap set of 74 genotypes. Within the A subgroup there is still clear population structure, but further subset selection would limit both the sample size and the phenotypic and genotypic diversity, reducing the rate of true positive results. This residual relatedness between lines was accounted for by including a kinship matrix in the GWAS model.

### Early vigor and ear emergence show genotypic variation in response to different simulated environments

To determine the genetic architecture for ear emergence date, early vigor, and a range of other agronomic traits (see *Materials and Methods*), the refined and balanced HapMap set of 74 *B. distachyon* accessions (File S10), with four biological replicates, were grown in two simulated conditions in climate chambers (Brown *et al*. 2014). To determine the effect of an increase in temperature in line with climate change predictions on the traits of interest, the conditions modeled a present (2015, Figure 2A) and a future (2050, Figure 2B) temperature profile at Wagga Wagga, NSW, Australia. The appropriate increase in average maximum and minimum temperature for each month were determined using an average of 12 global climate change models determined to be high confidence for south east Australia using the Climate Futures Tool (Figure 2C; Wilson *et al.* 2015a; File S11).

As expected, the accessions developed quicker and grew larger in the 2050 temperature profile (Figure 2, A and B) as is consistent with a quicker accumulation of thermal time (Figure 2D). Early vigor parameters and energy use efficiency traits were measured when the majority of plants were at a four-leaf stage. Growth stages, tiller numbers and ear emergence dates were monitored twice a week (File S12–S14). The experiment ceased after 200 days of growth, at which time there were five and seven lines that did not flower in the 2015 condition and 2050 conditions, respectively. The remaining lines reached ear emergence at a similar number of days in both the present and future conditions (Figure 2E). However, when converted to thermal time, those lines in the 2015 temperature condition required less thermal time than those in the 2050 temperature condition (Figure 2, D and F). This indicates that there is generally more dependence on photoperiod in this population than on thermal time to trigger the transition to flowering. There was variation between genotypes in the plasticity of their response to the two conditions (Figure 2, E and F), indicating that it would be worthwhile mapping the genotype by environment interaction ($G \times E$).

### Determining the genetic architecture of early growth, ear emergence, and energy use efficiency traits in response to environment

GWAS were performed on raw and derived traits as described in the *Materials and Methods* (Figure 3 and File S15 and File S16). All GWAS data are publicly available and traits are genome browseable online at https://easygwas.ethz.ch/gwas/myhistory/public/17/. For ear emergence, eight significant

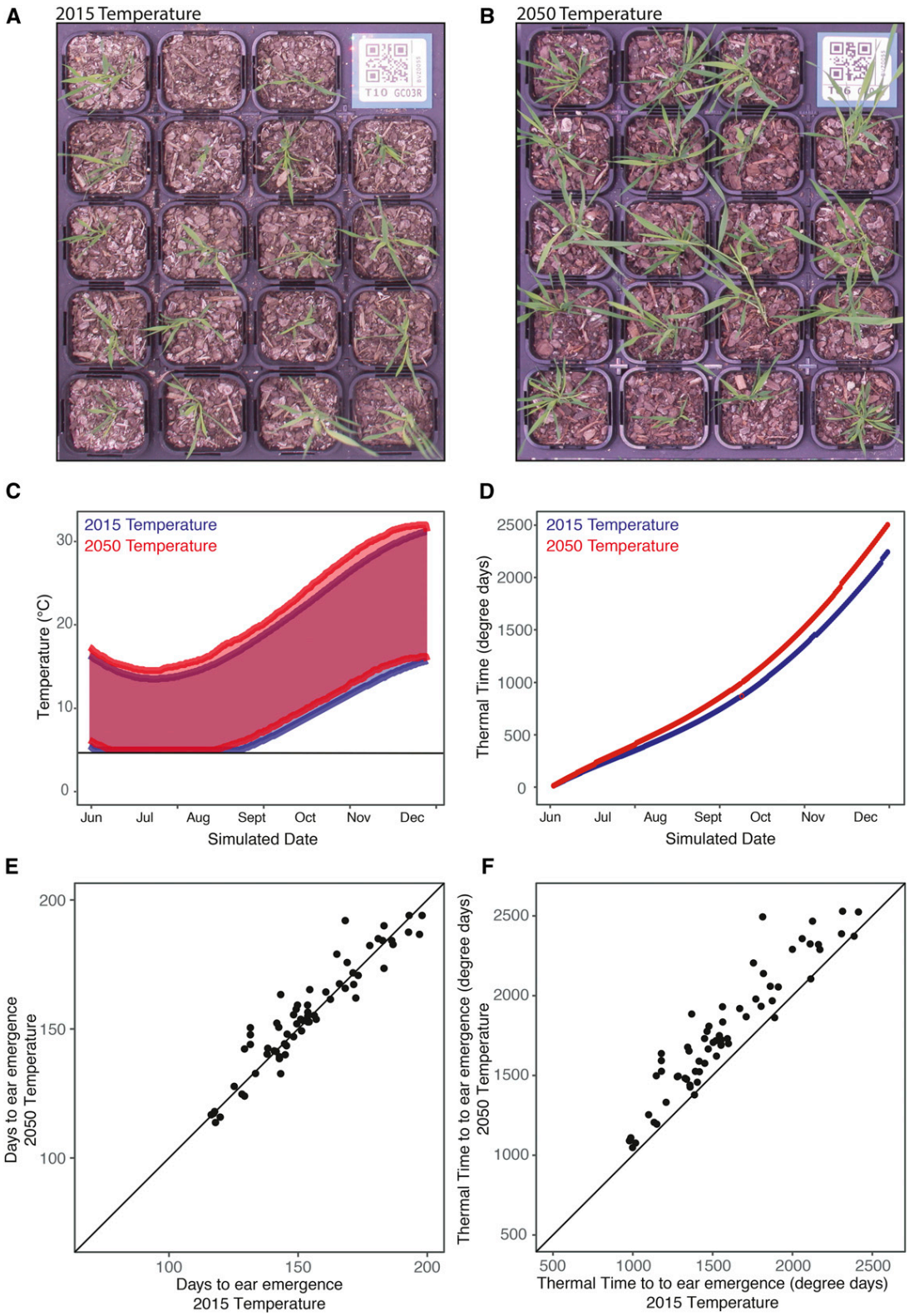



**Figure 2** Differential growth of *B. distachyon* under current and future climate growth temperatures. Climate chambers were used to compare the response of agronomic traits to small change in the climate for a Winter sowing in the Wagga Wagga region, south-eastern Australia. The GWAS HapMap set were grown in (A) 2015 temperature climate and (B) a 2050 temperature climate. Photos show representative plants after 16 weeks of growth. Climate chambers were programmed to have (C) diurnal and seasonal changes in temperature resulting in different rates of accumulation of thermal time (D) in the 2015 and 2050 climates. Timing of ear emergence was compared between chambers for both (E) days to ear emergence and (F) the accumulation of thermal time to ear emergence, demonstrating G × E interactions.

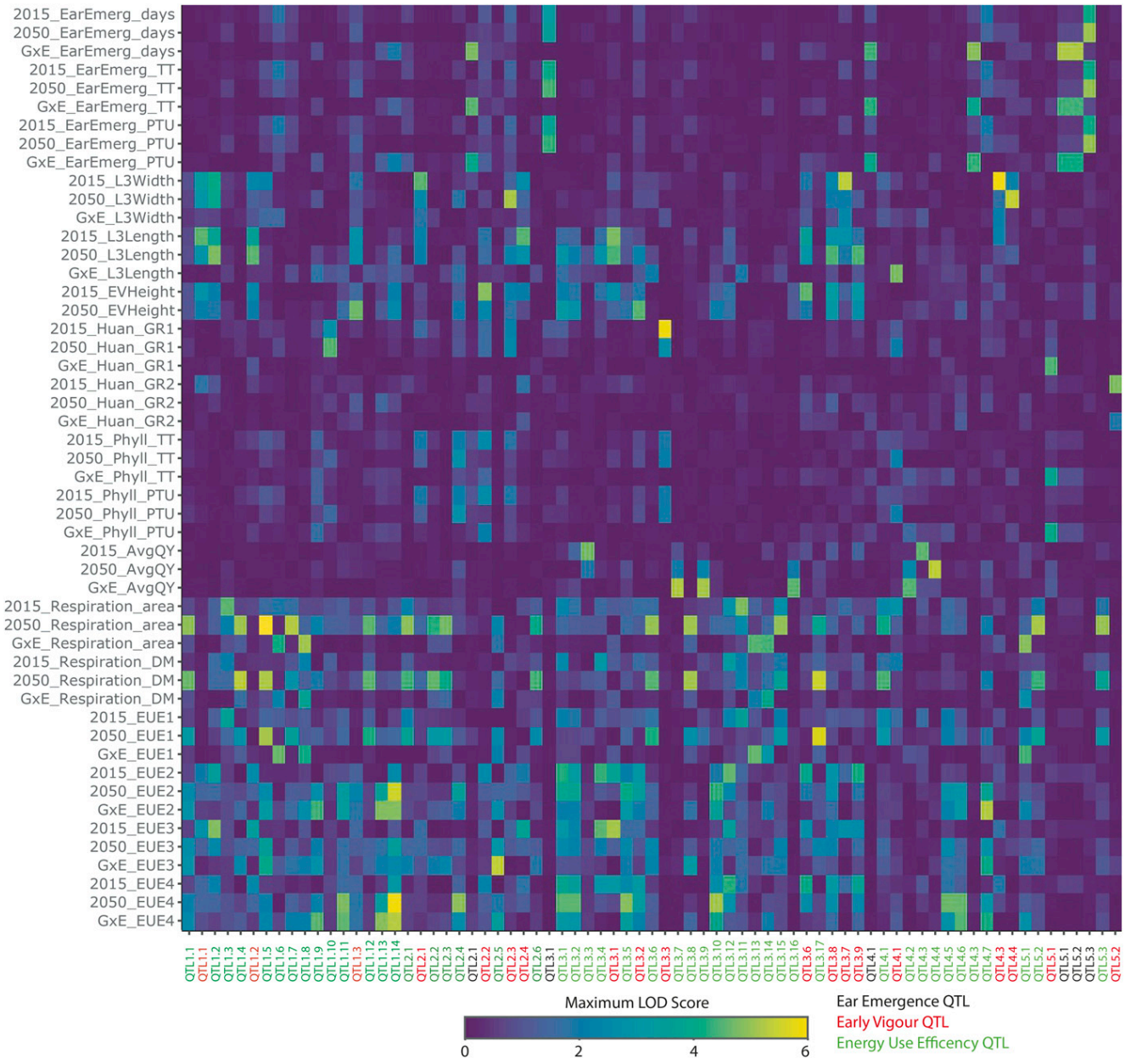



**Figure 3** Summary of QTL for each trait under current and future climate growth temperatures. A total of 73 significant QTL were identified for a range of agronomic traits phenotyped in the 2015 temperature and 2050 temperature climates and the G × E interaction. There was little overlap between QTL for different traits but two robust QTL were identified in both environments while 16 QTL were identified for a G × E interaction. G × E, genotype by environment interaction; EarEmerg, ear emergence; TT, thermal time; PTU, photothermal units; L3Width, leaf 3 width; L3Length, leaf 3 length; GR, growth rate; GR, growth rate; EV, early vigor; phyll, phyllacron interval; AvgQY, average quantum yield; DM, dry mass; EUE, energy use efficiency

QTL were identified. EarEmerg_QTL4.2 explains 62% of the phenotypic variation in thermal time to ear emergence in the 2015 temperature condition, while two QTL, EarEmerg_QTL3.1 and EarEmerg_QTL5.3, explain 56 and 10%, respectively, of the phenotypic variance in thermal time to ear emergence in the 2050 temperature condition. No QTL were found to be significant in both conditions but EarEmerg_QTL5.3 was significant in the 2050 temperature condition and was just under the significant threshold in the 2015 temperature condition (Figure 4A and File S17). Within the 100 kb region of this SNP there are 15 genes, several of which could be relevant to the regulation of flowering, including a YABBY transcription factor (Bradi5g16910), a no apical meristem (NAM) protein (Bradi5g16917), and an expressed gene containing a RNA recognition motif (Bradi5g16930). Interestingly, there were two QTL that were significant for thermal time to ear emergence, EarEmerg_QTL3.1 and EarEmerg_QTL4.2, but not for days to ear emergence. There were six QTL identified for the G × E interaction, explaining, in

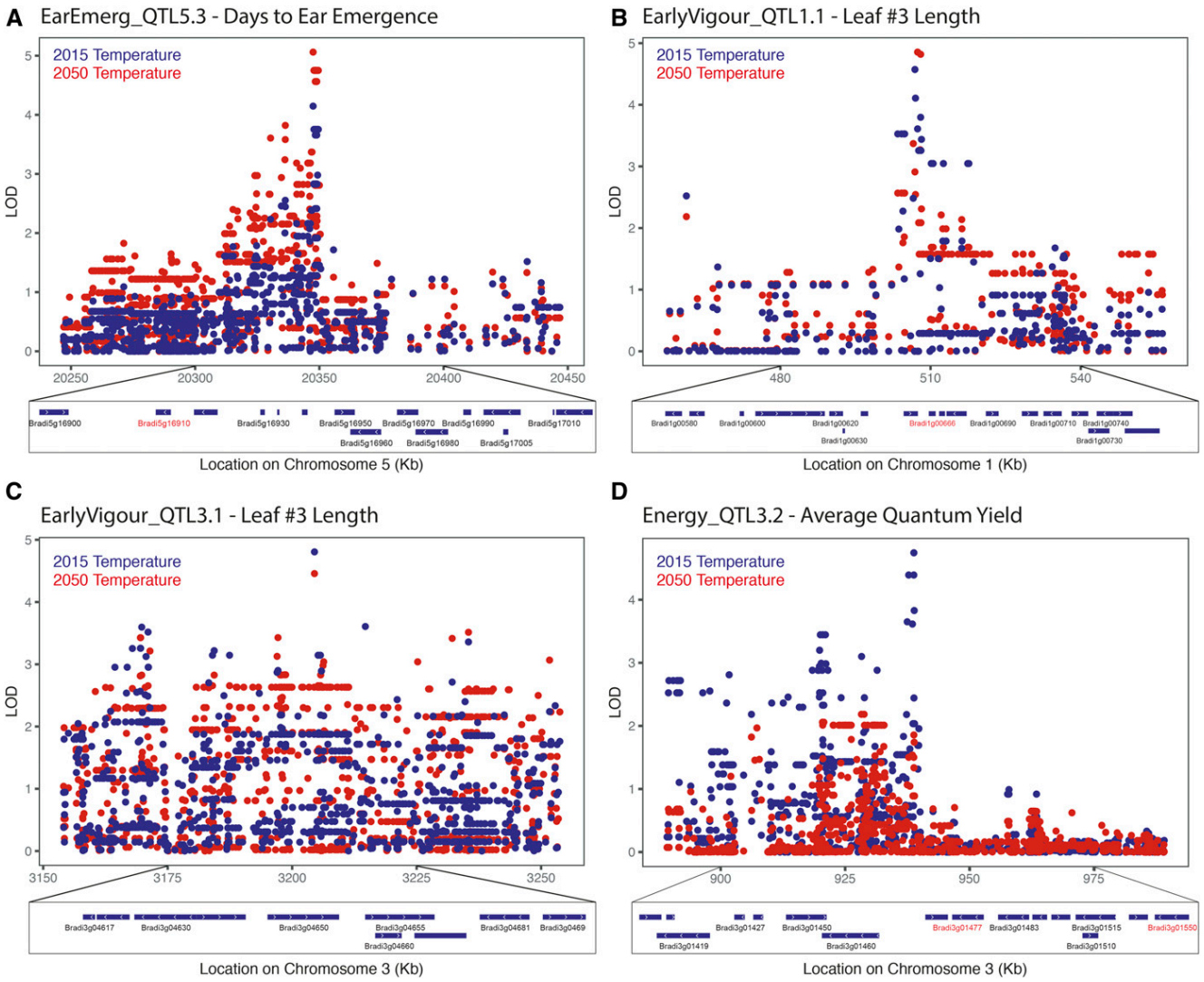



**Figure 4** Putative candidate genes for QTL of key interest. (A) The ear emergence QTL, EarEmerg_QTL5.3, was significant for days to ear emergence in the 2050 temperature condition and only just under the significance threshold for the 2015 condition. Likely candidate genes include a YABBY transcription factor Bradi5g16910. (B) The early vigor QTL, EarlyVigour_QTL1.1 for leaf #3 length was found to be significant in both conditions. This region contains an ethylene sensitive transcription factor, Bradi1g00666. (C) The early vigor QTL, EarlyVigour_QTL3.1 was also identified for leaf #3 length in both environments. (D) A strong QTL was identified for photosynthetic efficiency, Energy_QTL3.2, which was significant only in the 2015 temperature condition. Likely candidate genes include a heat shock protein, Bradi3g01477, and a Low PSII Accumulation 3 (LPA3) protein, Bradi3g01550. Locus identifiers in red represent these candidate genes.

part, the variation among lines in response to future climate.

For early vigor, 22 significant QTL were identified for five traits across the two climate conditions (File S15). Two QTL were identified in both conditions, EarlyVigour_QTL1.1 and EarlyVigour_QTL3.1, and both of these were for leaf #3 length. The 100-kb region surrounding these QTL contained 19 and 13 genes, respectively (Figure 4, B and C). There was a highly significant QTL on chromosome three for growth rate 1, EarlyVigour_QTL3.3, a measure of the rate of development of the seedling at the two leaf stage, but only in the 2015 temperature condition. The 100-kb region surrounding this QTL contained 13 genes (File S18). A total of six QTL were identified for the G × E interaction across the two conditions for early vigor traits.

For the energy use efficiency traits, a total of 47 QTL were identified across the two conditions for the three measured traits and four derived traits (File S15). Of these QTL, none were found in both environments. However, a strong QTL, Energy_QTL3.3, was identified for average quantum yield, a measure of photosynthetic efficiency, in the 2015 temperature environment. The 100-kb region around this QTL contained 24 genes including a low PSII accumulation three chloroplastic protein (Bradi3g01550), a Heat Shock Protein (Bradi3g01477) and several transcription factors (Figure 4D and File S19).

## Discussion

Thanks to the international *Brachypodium* community, in addition to our own collections, here we were able to provide the most comprehensive survey of *Brachypodium* species complex diversity to date. With 1968 accessions across the globe this is a >10-fold increase from previous studies (Filiz *et al.* 2009; Tyler *et al.* 2016).

Since being described as three separate species in 2012 (Catalán *et al.* 2012), species identification in the *Brachypodium* species complex has been achieved by morphology, PCR of a select set of markers or DNA barcoding (*e.g.*, Rebetzke *et al.* 1999a; Wilson *et al.* 2015b). Here, we present a unique systematic method of determining the species of an accession using low coverage GBS and bioinformatics, providing a high-throughput and low-cost alternative for species identification. We found that the majority of our accessions were *B. hybridum* (56%), including the vast majority of accessions in Australia and North America (Figure 1A). The wide dispersion of this species may be due to the benefit of the multiple genomes resulting from polyploidization (te Beest *et al.* 2012). There were relatively few *B. stacei* (3%), which were limited to the Mediterranean region (Figure 1A).

Within *B. distachyon* itself, we found significant population structure, including high level subgroup splits, with 6.5% of SNPs diverged between subgroups, which is greater than that found between *indica* and *japonica* rice at 1.4% divergence (Ma and Bennetzen 2004). While many previous studies have focused on individual regions (Filiz *et al.* 2009; Marques *et al.* 2017), the collection of 490 diverse *B. distachyon* accessions genotyped at 81,400 high quality SNPs presented here has allowed us to further distinguish groups with the *B. distachyon* subgroups, with an eastern and western European group in each subgroup. A number of geographically diverse highly related genotypic lineages were also identified, which showed within-lineage divergence of between 0.1 and 0.4% of SNPs. The geographic spread of these lineages highlights the inbreeding nature and high dispersal ability of *B. distachyon*.

The hierarchical levels of genetic variation within the *Brachypodium* species complex can be attributed to allopolyploidization and subspeciation, possibly during the most recent ice age; east/west IBD in Europe; and the high levels of self-fertilization in the species (Wilson *et al.* 2015a). These levels of population structure have been seen in Arabidopsis (Atwell *et al.* 2010), and other highly selfing plant species such as barley (Wang *et al.* 2012), but are more extreme in *Brachypodium*. In rice, either the *indica* sub-species (Huang *et al.* 2012) or *japonica* subspecies (Yano *et al.* 2016), have been separately used for GWAS. Similarly, to deal with the population structure in this study, the HapMap set was limited to the A subgroup of *B. distachyon* with remaining relatedness included in the GWAS analysis using mixed models (Cheng *et al.* 2011).

The lack of recombinant genetic diversity with subgroup and populations of *B. distachyon* also limits the power of GWAS analysis. The HapMap set contains a large amount of genomic diversity (>1% of bases are variable) but the sample size is low and the extent of LD is high, limiting mapping resolution. However, the patterns are similar to rice where GWAS is very effective as sample size increases (Huang and Han 2014). The construction of a Nested Association Mapping (NAM) population for *B. distachyon* would be advantageous to break-up the population and familial lineages, and to increase the frequency of minor alleles. This has been a successful approach in other species such as maize and wheat (Tian *et al.* 2011; Bajgain *et al.* 2016). Nevertheless, our set of lines and genomic data available in easyGWAS are an important resource for the community to map the genetic basis of various complex traits in this emerging model grass species. The small stature and rapid generation time of Brachypodium makes it especially useful for high throughput assays of phenomic traits across a range of controlled conditions.

In field conditions, determining the relationship between various physiological traits and their impact on yield is difficult due to seasonal environmental variability and the presence of a range of abiotic and biotic stresses. However, experiments in growth chambers often have little relevance to field conditions due to the unrealistic and static nature of the conditions. By using dynamic growth conditions, which mimic regional climates, we can avoid the stochastic downsides of field experiments while providing results arguably more translatable to the field (Brown *et al.* 2014; Poorter *et al.* 2016). The use of climate chambers also allows the impact of small changes in climate to be observed, and the dissection of which components of the climate have the largest influence on a trait of interest. In this study, we examined the effect of an increase in temperature in line with climate change model predictions for 2050 in south eastern Australia. Unexpectedly, there was generally a short delay of flowering time in the 2050 temperature condition, with variation in the extent of delay in different genotypes, while there was little dependence of flowering on the accumulation of thermal time. This suggests that there may be some vernalization requirements in *B. distachyon* that are not being met in the 2050 temperature condition. The lack of vernalization is also evident in the fact that seven lines had not flowered by the end of the 2050 temperature condition, while five lines did not flower in the 2015 temperature condition. While this GWAS analysis did not identify known flowering time loci that regulate vernalization-induced flowering such as VRN1, VRN2, and FT (Woods *et al.* 2014; Bettgenhaeuser *et al.* 2017), the QTL may represent more subtle vernalization processes that would be important for facultative varieties. Perhaps largely to the difference in growth conditions, the QTL in this study did not overlap with those found in a previous GWAS of flowering time (Tyler *et al.* 2016); this may also be an example of the Beavis effect (Xu 2003). Candidate genes identified for flowering time here included several transcription factors, including a YABBY transcription factor under EarEmerg_QTL5.3. The closest rice ortholog, Os04g45330,

to this YABBY transcription factor is most highly expressed in the shoot apical meristem and developing inflorescence (Rice Gene Expression Atlas), while the closest Arabidopsis ortholog, At2g45190, is involved in regulation of the floral morphology (Lu *et al.* 2007). This EarEmerg_QTL5.3 was significant in the 2050 temperature conditions and was only just below the significance threshold in the 2015 temperature condition (Figure 4A).

Early vigor is an important trait in many parts of Australia and the rest of the world, where there is competition from weeds and a shorter season. Despite the highest correlating nondestructive measure of early above ground biomass being seedling height, the most robust QTL across environments were actually identified by leaf #3 length. Two QTL identified for leaf #3 length were identified in both environments, indicating they could potentially be useful for breeding for early vigor in multiple environment types. One of these, EarlyVigour_QTL1.1 is located in an area of synteny to other areas where early vigor QTL have been identified at the end of chromosome 3 in rice (Lu *et al*. 2007; The International Brachypodium Initiative 2009; Singh *et al*. 2017) and Chromosome 4 in wheat (Rebetzke *et al.* 2001). Within EarlyVigour_QTL1.1 there is a candidate gene, Bradi1g00666, that is described as an ethylene-responsive transcription factor. The main candidate gene in the QTL on chromosome 3 in rice was also an ethylene responsive gene (Singh *et al.* 2017). The EarlyVigour_QTL3.1 for leaf #3 length was also found to be significant across both environments. There were no obvious candidate genes for this QTL, but a number of signaling proteins that could be involved in molecular control of leaf size (Figure 4C and File S18).

The balance of energy production and use in plants is highly linked to the conditions that the plant is grown under; however, genetic variation controlling the energy efficiency of plants could be used to increase yield potentials. The quantum yield is an indicator of photosynthetic efficiency, the proportion of energy harvested through the light-harvesting complexes that goes toward producing photosynthates (Rungrat *et al.* 2016). No QTL were identified in common across both environments, but there were 11 QTL that were identified for the G $\times$ E interaction. This may be due to the sensitivity of these energy processes to the subtle difference in environments or a result of being measured on different days to allow comparison of plants at the same developmental stage. A strong QTL was identified for quantum yield, a measure of the efficiency of Photosystem II (PSII), in the 2015 climate but, interestingly, not in the 2050 climate. Candidate genes under this QTL included a gene with 66% homology to the Low PSII Accumulation 3 (LPA3) gene in Arabidopsis, which has been shown to be important in PSII assembly (Lu 2016). Further studies into the importance of this QTL in different conditions, as well as the other photosynthesis and respiration QTL, would be worthwhile.

In conclusion, the *Brachypodium* species complex is heavily structured at the ploidy, subgroups, population, and family levels. This limits the ability to identify the genetic basis of adaptation as relatively few recombinant genotypes were obtained. Despite these limitations, this study indicates the potential to use *Brachypodium distachyon*, a model for Pooideae grass crops, to identify genetic variation in key pathways underlying agricultural traits through GWAS. Further wild collections and/or the development of NAM populations could address the limitation of recombinant genotypes and result in very high power mapping population typical of 1000 genome projects. As it now stands, *Brachypodium* is a good model for both polyploidization, with likely multiple events among small divergent genomes, and for invasion biology with multiple widespread genotypes identified across continents, regions, and sites.

## Acknowledgments

Brachypodium accession contributors (PI: principle investigator): ShuangShuang Liu, Kent Bradford (PI), Smadar Ezrati (PI), Hikmet Budak (PI), Diana Lopez, Pilar Catalan (PI), David Garvin (PI), John Vogel (PI), Sean Gordon, Sam Hazen (PI), Luis Mur (PI). We would like to acknowledge the technical assistance of Suyan Yee and Allison Heussler. We thank Scott Ferguson and Dominik Grimm for their help adding SNPs and traits into easyGWAS. We are grateful for funding and support from the Australian Research Council Centre of Excellence in Plant Energy Biology (CE140100008). Australian Plant Phenomics Facility is supported under the National Collaborative Research Infrastructure Strategy of the Australian Government. The research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

Author contributions: P.B.W., J.C.S., N.C.A., N.W., and S.P.G. undertook the experiments; P.B.W., J.C.S., J.P.V., and J.O.B. designed the study concept and experiments; P.B.W., J.C.S., K.D.M., S.R.E., R.C., K.S., and S.P.G. undertook the analysis of the data; P.B.W., J.C.S., K.D.M., S.R.E., and J.O.B. wrote the manuscript, and all authors read the manuscript.

## Literature Cited

Atwell, S., Y. S. Huang, B. J. Vilhjalmsson, G. Willems, M. Horton *et al.*, 2010 Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465: 627–631. https://doi.org/10.1038/nature08800

Bajgain, P., M. N. Rouse, T. J. Tsilo, G. K. Macharia, S. Bhavani *et al.*, 2016 Nested association mapping of stem rust resistance in wheat using genotyping by sequencing. PLoS One 11: e0155760. https://doi.org/10.1371/journal.pone.0155760

Bettgenhaeuser, J., F. M. K. Corke, M. Opanowicz, P. Green, I. Hernández-Pinzón *et al.*, 2017 Natural variation in *Brachypodium* links vernalization and flowering time loci as major flowering determinants. Plant Physiol. 173: 256–268. Available at: http://www.plantphysiol.org/lookup/doi/10.1104/pp.16.00813. https://doi.org/10.1104/pp.16.00813

Brachi, B., N. Faure, M. Horton, E. Flahauw, A. Vazquez *et al.*, 2010 Linkage and association mapping of Arabidopsis thaliana flowering time in nature. PLoS Genet. 6: e1000940. https://doi.org/10.1371/journal.pgen.1000940

Brachi, B., G. P. Morris, and J. O. Borevitz, 2011 Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.* 12: 232. https://doi.org/10.1186/gb-2011-12-10-232

Brown, T. B., R. Cheng, X. R. R. Sirault, T. Rungrat, K. D. Murray *et al.*, 2014 TraitCapture: genomic and environment modelling of plant phenomic data. Curr. Opin. Plant Biol. 18: 73–79. Available at: http://www.sciencedirect.com/science/article/pii/S1369526614000181. https://doi.org/10.1016/j.pbi.2014.02.002

Brutnell, T. P., J. L. Bennetzen, and J. P. Vogel, 2015 *Brachypodium* distachyon and Setaria viridis: model genetic systems for the grasses. Annu. Rev. Plant Biol. 66: 465–485. https://doi.org/10.1146/annurev-arplant-042811-105528

Catalán, P., J. Muller, R. Hasterok, G. Jenkins, L. A. J. Mur *et al.*, 2012 Evolution and taxonomic split of the model grass *Brachypodium* distachyon. Ann. Bot. 109: 385–405. https://doi.org/10.1093/aob/mcr294

Catalán, P., D. Lopez-Alvarez, C. Bellosta, and L. Villar, 2016 Updated taxonomic descriptions, iconography, and habitat preferences of *Brachypodium* distachyon, B. stacei, and B. hybridum (Poaceae). An del Jard Bot Madrid. 73: e028. https://doi.org/10.3989/ajbm.2428

Cheng, R., M. Abney, A. A. Palmer, and A. D. Skol, 2011 QTLRel: an R package for genome-wide association studies in which relatedness is a concern. BMC Genet. 12: 66. https://doi.org/10.1186/1471-2156-12-66

Climate Change in Australia website, 2015 Available at: https://www.climatechangeinaustralia.gov.au/en/climate-projections/climate-futures-tool/introduction-climate-futures/. Accessed: September 1, 2015

Condon, A. G., R. A. Richards, G. J. Rebetzke, and G. D. Farquhar, 2004 Breeding for high water-use efficiency. J. Exp. Bot. 55: 2447–2460. https://doi.org/10.1093/jxb/erh277

Danecek, P., S. Schiffels, and R. Durbin, 2016 Multiallelic calling model in bcftools (-m). Available at: http://samtools.github.io/bcftools/call-m.pdf.

Draper, J., L. A. J. Mur, G. Jenkins, G. C. Ghosh-Biswas, P. Bablak *et al.*, 2001 *Brachypodium* distachyon. A new model system for functional genomics in grasses. Plant Physiol. 127: 1539–1555. Available at: http://www.plantphysiol.org/content/127/4/1539. https://doi.org/10.1104/pp.010196

Earl, D. A., and B. M. vonHoldt, 2012 STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 4: 359–361. https://doi.org/10.1007/s12686-011-9548-7

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379. https://doi.org/10.1371/journal.pone.0019379

Evanno, G., S. Regnaut, and J. Goudet, 2005 Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14: 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Filiz, E., B. S. Ozdemir, F. Budak, J. P. Vogel, M. Tuna *et al.*, 2009 Molecular, morphological, and cytological analysis of diverse *Brachypodium* distachyon inbred lines. Genome 52: 876–890. https://doi.org/10.1139/G09-062

Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907v2 [q-bio.GN]

Garvin, D. F., Y. Q. Gu, R. Hasterok, S. P. Hazen, G. Jenkins *et al.*, 2008 Development of genetic and genomic research resources for *Brachypodium* distachyon, a new model system for grass crop research. Crop Sci. 48: 69–84.

Giraldo, P., M. Rodríguez-Quijano, J. F. Vázquez, J. M. Carrillo, and E. Benavente, 2012 Validation of microsatellite markers for cytotype discrimination in the model grass *Brachypodium* distachyon. Genome 55: 523–527. https://doi.org/10.1139/g2012-039

Gordon, S. P., B. Contreras-Moreira, D. P. Woods, and J. P. Vogel, 2017 Extensive gene content variation in the *Brachypodium* distachyon pan-genome correlates with population structure. Nat. Commun. 8: 2184. https://doi.org/10.1038/s41467-017-02292-8

Hasterok, R., J. Draper, and G. Jenkins, 2004 Laying the cytotaxonomic foundations of a new model grass, *Brachypodium* distachyon (L.) beauv. Chromosome Res. 12: 397–403. https://doi.org/10.1023/B:CHRO.0000034130.35983.99

Haun, J. R., 1973 Visual quantification of wheat development. Agron. J. 65: 116–119. https://doi.org/10.2134/agronj1973.00021962006500010035x

Higgins, J. A., P. C. Bailey, and D. A. Laurie, 2010 Comparative genomics of flowering time pathways using *Brachypodium* distachyon as a model for the temperate grasses. PLoS One 5: e10065. https://doi.org/10.1371/journal.pone.0010065

Huang, X., and B. Han, 2014 Natural variations and genome-wide association studies in crop plants. Annu. Rev. Plant Biol. 65: 531–551. https://doi.org/10.1146/annurev-arplant-050213-035715

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng *et al.*, 2010 Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42: 961–967. https://doi.org/10.1038/ng.695

Huang, X., Y. Zhao, X. Wei, C. Li, A. Wang *et al.*, 2012 Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat. Genet. 44: 32–39. https://doi.org/10.1038/ng.1018

Idziak, D., I. Hazuka, B. Poliwczak, A. Wiszynska, E. Wolny *et al.*, 2014 Insight into the karyotype evolution of *Brachypodium* species using comparative chromosome barcoding. PLoS One 9: e93503. https://doi.org/10.1371/journal.pone.0093503

Jakobsson, M., and N. A. Rosenberg, 2007 CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801–1806. https://doi.org/10.1093/bioinformatics/btm233

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN].

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Listgarten, J., C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin *et al.*, 2012 Improved linear mixed models for genome-wide association studies. Nat. Methods 9: 525–526. https://doi.org/10.1038/nmeth.2037

López-Alvarez, D., M. L. López-Herranz, A. Betekhtin, and P. Catalán, 2012 A DNA barcoding method to discriminate between the model plant *Brachypodium distachyon* and its close relatives *B. stacei* and *B. hybridum* (Poaceae). PLoS One 7: e51058. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3519806&tool=pmcentrez&rendertype=abstract. https://doi.org/10.1371/journal.pone.0051058

Lu, X. L., A. L. Niu, H. Y. Cai, Y. Zhao, J. W. Liu *et al.*, 2007 Genetic dissection of seedling and early vigor in a recombinant inbred line population of rice. Plant Sci. 172: 212–220. https://doi.org/10.1016/j.plantsci.2006.08.012

Lu, Y., 2016 Identification and roles of photosystem II assembly, stability, and repair factors in Arabidopsis. Front. Plant Sci. 7: 168. Available at: http://journal.frontiersin.org/Article/10.3389/fpls.2016.00168/abstract. https://doi.org/10.3389/fpls.2016.00168

Ma, J., and J. L. Bennetzen, 2004 Rapid recent growth and divergence of rice nuclear genomes. Proc. Natl. Acad. Sci. USA 101: 12404–12410. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.0403715101. https://doi.org/10.1073/pnas.0403715101

Marques, I., V. Shiposha, D. López-Alvarez, A. J. Manzaneda, P. Hernandez *et al.*, 2017 Environmental isolation explains Iberian genetic diversity in the highly homozygous model grass *Brachypodium* distachyon. BMC Evol. Biol. 17: 139. https://doi.org/10.1186/s12862-017-0996-x

Maurer, A., V. Draba, Y. Jiang, F. Schnaithmann, R. Sharma *et al.*, 2015 Modelling the genetic architecture of flowering time control in barley through nested association mapping. BMC Genomics 16: 290. Available at: http://www.biomedcentral.com/1471–2164/16/290. https://doi.org/10.1186/s12864-015-1459-7

Morris, G. P., P. P. Grabowski, and J. O. Borevitz, 2011 Genomic diversity in switchgrass (*Panicum virgatum*): from the continental scale to a dune landscape. Mol. Ecol. 20: 4938–4952. https://doi.org/10.1111/j.1365-294X.2011.05335.x

Mur, L. A. J., J. Allainguillaume, P. Catalan, R. Hasterok, G. Jenkins *et al.*, 2011 Exploiting the *Brachypodium* tool box in cereal and grass research. New Phytol. 191: 334–347. https://doi.org/10.1111/j.1469-8137.2011.03748.x

Murray, K. D., and J. O. Borevitz, 2017 libqcpp: a C++14 sequence quality control library. J Open Source Softw. 2: 232. https://doi.org/10.21105/joss.00232

Murray, K. D., and J. O. Borevitz, 2018 Axe: rapid, competitive sequence read demultiplexing using a trie. Bioinformatics 34: 3924–3925. https://doi.org/10.1093/bioinformatics/bty432

Nicotra, A. B., C. Chong, J. G. Bragg, C. R. Ong, N. C. Aitken *et al.*, 2016 Population and phylogenomic decomposition via genotyping-by-sequencing in Australian Pelargonium. Mol. Ecol. 25: 2000–2014. https://doi.org/10.1111/mec.13584

Pedersen, B. S., and A. R. Quinlan, 2017 Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. Available at: http://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btx699/4583630.

Poorter, H., F. Fiorani, R. Pieruschka, T. Wojciechowski, W. H. van der Putten *et al.*, 2016 Pampered inside, pestered outside? Differences and similarities between plants growing in controlled conditions and in the field. New Phytol. 212: 838–855. https://doi.org/10.1111/nph.14243

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Ray, D. K., N. D. Mueller, P. C. West, and J. A. Foley, 2013 Yield trends are insufficient to double global crop production by 2050. PLoS One 8: e66428. https://doi.org/10.1371/journal.pone.0066428

Ream, T. S., D. P. Woods, C. J. Schwartz, C. P. Sanabria, J. Mahoy *et al.*, 2014 Interaction of photoperiod and vernalization determines flowering time of *Brachypodium* distachyon. Plant Physiol. 164: 694–709. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3912099&tool=pmcentrez&rendertype=abstract. https://doi.org/10.1104/pp.113.232678

Rebetzke, G. J., and R. A. Richards, 1999 Genetic improvement of early vigour in wheat. Aust. J. Agric. Res. 50: 291–301. https://doi.org/10.1071/A98125

Rebetzke, G. J., R. Appels, A. D. Morrison, R. A. Richards, G. McDonald *et al.*, 2001 Quantitative trait loci on chromosome 4B for coleoptile length and early vigour in wheat (*Triticum aestivum* L.). Aust. J. Agric. Res. 52: 1221–1234. https://doi.org/10.1071/AR01042

Rivers, J., N. Warthmann, B. J. Pogson, and J. O. Borevitz, 2015 Genomic breeding for food, environment and livelihoods. Food Secur. 7: 375–382. https://doi.org/10.1007/s12571-015-0431-3

Rungrat, T., M. Awlia, T. Brown, R. Cheng, X. Sirault *et al.*, 2016 Using phenomic analysis of photosynthetic function for abiotic stress response gene discovery. Arabidopsis Book 14: e0185.

Scafaro, A. P., A. C. A. Negrini, B. O'Leary, F. A. A. Rashid, L. Hayes *et al.*, 2017 The combination of gas-phase fluorophore technology and automation to enable high-throughput analysis of plant respiration. Plant Methods 13: 16. https://doi.org/10.1186/s13007-017-0169-3

Schubert, M., S. Lindgreen, and L. Orlando, 2016 AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes 9: 88. https://doi.org/10.1186/s13104-016-1900-2

Singh, U. M., S. Yadav, S. Dixit, P. J. Ramayya, M. N. Devi *et al.*, 2017 QTL hotspots for early vigor and related traits under dry direct-seeded system in rice (Oryza sativa L.). Front. Plant Sci. 8: 286. Available at: https://www.frontiersin.org/article/10.3389/fpls.2017.00286. https://doi.org/10.3389/fpls.2017.00286

Spokas, K., and F. Forcella, 2006 Estimating hourly incoming solar radiation from limited meteorological data. Weed Sci. 54: 182–189. Available at: https://www.cambridge.org/core/product/identifier/S0043174500007670/type/journal_article. https://doi.org/10.1614/WS-05-098R.1

Sukumaran, S., S. Dreisigacker, M. Lopes, P. Chavez, and M. P. Reynolds, 2014 Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. Theor. Appl. Genet. 128: 353–363. https://doi.org/10.1007/s00122-014-2435-3

te Beest, M., J. J. Le Roux, D. M. Richardson, A. K. Brysting, J. Suda *et al.*, 2012 The more the better? The role of polyploidy in facilitating plant invasions. Ann. Bot. 109: 19–45. https://doi.org/10.1093/aob/mcr277

The International Barley Genome Sequencing Consortium, 2012 A physical, genetic and functional sequence assembly of the barley genome. Nature 491: 711–716. https://doi.org/10.1038/nature11543

The International Brachypodium Initiative, 2009 Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463: 763–768. Available at: http://link.springer.com/10.1007/s11103–009–9456–3.

The International Wheat Genome Sequencing Consortium, 2017. Available at: www.wheatgenome.org.

Tian, F., P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun *et al.*, 2011 Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. 43: 159–162. https://doi.org/10.1038/ng.746

Tyler, L., S. J. Lee, N. D. Young, G. A. DeIulio, E. Benavente *et al.*, 2016 Population structure in the model grass is highly correlated with flowering differences across broad geographic areas. Plant Genome 9: doi: 10.3835/plantgenome2015.08.0074.

Vogel, J. P., M. Tuna, H. Budak, N. Huo, Y. Q. Gu *et al.*, 2009 Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium* distachyon. BMC Plant Biol. 9: 88. https://doi.org/10.1186/1471-2229-9-88

Wang, M., N. Jiang, T. Jia, L. Leach, J. Cockram *et al.*, 2012 Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. Theor. Appl. Genet. 124: 233–246. https://doi.org/10.1007/s00122-011-1697-2

Wheeler, T., and J. von Braun, 2013 Climate change impacts on global food security. Science 341: 508–513. Available at: http://science.sciencemag.org/content/341/6145/508.abstract.

Wicker, T., K. F. X. Mayer, H. Gundlach, M. Martis, B. Steuernagel *et al.*, 2011 Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. Plant Cell 23: 1706–1718. Available at: http://www.plantcell.org/lookup/doi/10.1105/tpc.111.086629. https://doi.org/10.1105/tpc.111.086629

Wilson, D., and J. Jones, 1982 Effect of selection for dark respiration rate of mature leaves on crop yields of Lolium perenne cv. S23. Ann. Bot. 49: 313–320. Available at: http://aob.oxfordjournals.org/content/49/3/313.short. https://doi.org/10.1093/oxfordjournals.aob.a086255

Wilson, P. B., J. C. Streich, and J. O. Borevitz, 2015a Genomic diversity and climate adaptation in *Brachypodium*, pp. 107–127 in *Genetics and Genomics of Brachypodium*, edited by J. Vogel. Springer International Publishing, Cham, Switzerland. https://doi.org/10.1007/7397_2015_18

Wilson, P. B., G. J. Rebetzke, and A. G. Condon, 2015b Of growing importance : combining greater early vigour and transpiration efficiency for wheat in variable rainfed environments. Funct. Plant Biol. 42: 1107–1115.

Wilson, P. B., G. J. Rebetzke, and A. G. Condon, 2015c Pyramiding greater early vigour and integrated transpiration efficiency in bread wheat; trade-offs and benefits. F. Crop. Res. 183: 102–110. https://doi.org/10.1016/j.fcr.2015.07.002

Winzeler, M., D. E. McCullough, and L. A. Hunt, 1988 Genotypic differences in dark respiration of mature leaves in winter wheat (Triticum aestivum L.). Can. J. Plant Sci. 68: 669–675. https://doi.org/10.4141/cjps88-080

Woods, D. P., T. S. Ream, and R. M. Amasino, 2014 Memory of the vernalized state in plants including the model grass *Brachypodium* distachyon. Front. Plant Sci. 5: 99. https://doi.org/10.3389/fpls.2014.00099

Woods, D. P., R. Bednarek, F. Bouché, S. P. Gordon, J. P. Vogel *et al.*, 2017 Genetic architecture of flowering-time variation in *Brachypodium* distachyon. Plant Physiol. 173: 269–279. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27742753%0Ahttp://www.plantphysiol.org/lookup/doi/10.1104/pp.16.01178. https://doi.org/10.1104/pp.16.01178

Xu, S., 2003 Theoretical basis of the Beavis effect. Genetics 165: 2259–2268.

Yano, K., E. Yamamoto, K. Aya, H. Takeuchi, P. C. Lo *et al.*, 2016 Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat. Genet. 48: 927–934. Available at: http://www.nature.com/doifinder/10.1038/ng.3596. https://doi.org/10.1038/ng.3596

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie *et al.*, 2012 A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28: 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

*Communicating editor: M. Johnston*

# Chapter 6

# Landscape drivers of genomic diversity and divergence in woodland eucalypts

This chapter describes my recent work on the landscape drivers of genetic diversity in two species of woodland eucalypt, *Eucalyptus albens* and *Eucalyptus sideroxylon*. We found high genetic diversity, low differentiation between localities, no strong discrete population structure, and moderate to strong isolation by distance and environment. I performed and interpreted the analyses I present here (including creating bespoke analysis software and pipelines), and wrote the manuscript. My co-authors contributed significantly to sample collection, creation of sequencing libraries, and interpretation.

This work will soon be submitted to Molecular Ecology and biorXiv, and will be presented at the Eucalyptus Genomics 2019 conference. The author list will be: Kevin Murray, Jasmine Janes, Helen Bothwell, Ashley Jones, Rose Andrew, Justin Borevitz.

## 6.1 Abstract

Spatial genetic patterns are influenced by numerous factors, and they can vary even among coexisting, closely related species due to differences in dispersal and selection. Eucalyptus (L'Héritier 1789; the "eucalypts") are foundation tree species that provide essential habitat and modulate ecosystem services throughout Australia. Here we present a study of landscape genomic variation in two woodland eucalypt species, using whole genome sequencing of 388 individuals of *Eucalyptus albens* and *Eucalyptus sideroxylon*. We found exceptionally high genetic diversity ($\pi \approx 0.05$) and low genome-wide, inter-specific differentiation ($F_{ST} = 0.15$). We found no support for strong, discrete population structure, but found substantial support for isolation by geographic distance (IBD) in both species. Using generalised dissimilarity modelling, we identified additional isolation by environment (IBE). *Eucalyptus albens* showed moderate IBD, and environmental variables have a small but significant amount of additional predictive power (i.e., IBE). *Eucalyptus sideroxylon* showed much stronger IBD, and moderate IBE. These results highlight the vast adaptive potential of these species, and set the stage for testing evolutionary hypotheses of interspecific adaptive differentiation across environments.

## 6.2 Introduction

In wild species, and especially plants, genetic variation is inherently spatial: individuals occur at specific locations, and allele frequencies differ across the landscape as a result of variation in demographic history, patterns of gene flow, and heterogeneous selection pressures. Landscape genomics is the study of the geographic distribution of alleles within a species and the underlying processes that shape gene flow. By interrogating spatial genetic patterns, we may examine the historical drivers of local genetic isolation and potential adaptation, and use this knowledge to better manage species under a changing environment (Hoffmann et al., 2015).

A multitude of processes may drive the spatial patterns of genetic diversity within and between species. Individuals may cluster into discrete genetic groups, with reduced gene flow between subpopulations relative to within. There are many potential causes of such discrete structure, for example geographic barriers to gene flow or flowering time divergence. Individuals may also exhibit patterns of continuous isolation by geographic distance (IBD; Wright, 1943) or isolation by environment (IBE; Wang and Bradburd, 2014). IBD is indicated by a positive correlation between increasing genetic dissimilarity and geographic distance, and is observed when individuals are more likely to reproduce with geographically-proximate individuals. IBE is indicated by a correlation between genetic dissimilarity and environmental

dissimilarity, while controlling for IBD. IBE can have many causes, for example environmental effects on phenology altering flowering time, or impeded dispersal between habitats due to maladaptation to local conditions. Any of these three patterns of genetic isolation over the landscape (discrete structure, IBD, or IBE) may occur within a given species. Importantly, these patterns describe genome-wide phenomena, and while they may be influenced or initially generated by selection on adaptive alleles, their detection is not evidence of local adaptation. While factors affecting dispersal, such as landscape resistance (Spear et al., 2010; Wang and Bradburd, 2014; Zeller et al., 2012), may vary across the landscape, much can be learned by applying these global, homogeneous, dissimilarity-based methods for studying IBD and IBE, particularly when integrated with tests of discrete genetic structure.

The processes that influence spatial autocorrelation of allele frequencies require sophisticated statistical methods to disentangle. Continuous isolation by distance can lead to support for discrete population structure in analysis with genetic clustering methods like STRUCTURE and ADMIXTURE (Frantz et al., 2009). However, recent methodological developments now allow joint estimation of IBD and discrete structure (conStruct; Bradburd et al., 2017). Spatial autocorrelation of environmental variables makes disentangling their effects from IBD challenging, and older methods like partial Mantel tests are beset with several flaws (e.g., assumption of linearity, high Type I error rate; Guillot and Rousset, 2013). Generalised dissimilarity modelling (GDM; Ferrier et al., 2002, 2007) is a method which can accurately discriminate the geographic and environmental contributions to genetic differentiation, even where effects are non-linear. Equally important is the selection of variables appropriate to one's study system: Williams et al. (2012) propose a comprehensive variable set and variable selection methodology specifically for ecological models of habitats. However sophisticated the methods used to detect isolation by environment, it is a pattern affecting the genomic background. Locally adaptive loci should stand out above this background and could be identified subsequently via a genome scan.

Genus *Eucalyptus* (L'Héritier; the "eucalypts") is a speciose lineage of trees and large shrubs that includes the keystone species of many Australian habitats. Box-gum grassy woodlands are one such habitat, and while once common in southeastern Australia, their conversion to agricultural land has reduced their range significantly (NSW Scientific Committee, 2002). We sought to examine spatial genetic patterns in two foundation species of these grassy woodlands, *Eucalyptus albens* (Benth.; "white box") and *Eucalyptus sideroxylon* (A. Cunn. ex Wools; "mugga ironbark"). The prevalence of discrete population structure, IBD and/or IBE has been studied in several eucalypt species (e.g. Andrew et al., 2005, 2007; Jones et al.,

2007; Jordan et al., 2017; Rutherford et al., 2018; Steane et al., 2006, 2015, 2014; Supple et al., 2018). Although eucalypts have very limited seed dispersal, they generally preferentially outcross and are pollinated by generalist bird and insect pollinators, both of which contribute to their spatial genetic structure (Booth, 2017; Potts and Gore, 1995; Williams and Woinarski, 1997). Spatial genetic autocorrelation is strong within populations, but tends to be weak at larger scales; for example, isolation by distance between localities is only apparent between localities separated by more than 500 km in E. melliodora (Supple et al., 2018). While many studies have tested for and found discrete genetic structure (e.g. in *E. globulus*; Steane et al., 2006), strong discrete genetic structure uncorrelated with geography has been reported less commonly in widespread eucalypt species (e.g. in *E. salubris*; Steane et al., 2015). In any case, given the likely conflation of IBD and discrete population structure by traditional genetic clustering methods (Bradburd et al., 2017; Frantz et al., 2009), the relative extent of IBD and discrete structure remains an open question in many species. Correlation between genetic variation and environment has been observed in many forms, including IBE (e.g. Supple et al., 2018) and genotype-environment associations (e.g. Jordan et al., 2017; Dillon et al., 2014; Steane et al., 2017a, 2017b, 2014).

We aimed to determine the relative influence of the various factors contributing to landscape-scale spatial genetic patterns in *E. albens* and *E. sideroxylon*. The large estimated census sizes (González-Orozco et al., 2016) of both species led us to predict that these species would exhibit high genetic diversity. The reproductive ecology and extensive latitudinal geographic ranges of these species, and previous results for closely related species, led us to expect weak patterns of IBD and little discrete population structure orthogonal to IBD in both these species. Given gene-environment associations observed in closely-related species, we also predicted that isolation by environment would be observed, particularly associations between genetic distance and variables describing the availability of and demand for moisture and nutrients. To test these hypotheses, we generated whole-genome sequence data for 215 and 173 individuals of *E. albens* and *E. sideroxylon*, respectively. We quantified intraspecific genetic variation across the landscape, determined the extent of both continuous isolation by distance and isolation by environment, and assessed discrete population structure independent of IBD.

## 6.3 Methods

### 6.3.1 Study system

The genus *Eucalyptus* (L'Héritier 1789; the "eucalypts") is described as a highly speciose lineage of trees and large shrubs within family *Myrtaceae*. Of the more than 800 described species (Nicolle, 2018; Pryor and Johnson, 1971) that have evolved over the last 70 My (Thornhill et al., 2015), nearly all are endemic to the Australian continent, with a small number of species occurring in Indonesia and New Guinea. Here we focus on two woodland eucalypt species. *Eucalyptus albens* and *E. sideroxylon* are from different series (*Buxeales* and *Melliodorae*, respectively) within *Eucalyptus* section *Adnataria*. They are morphologically distinct, differing in bark type (box vs ironbark) and flower size and colour (*E. sideroxylon* larger, sometimes pink-red pigmented; Brooker and Kleinig, 2006; Boland et al., 2006; Costermans, 1983). Both generally occur inland of the Great Dividing Range, with *E. sideroxylon*'s range extending further inland, while *E. albens* extends further south and has disjunct populations in southeast Victoria and South Australia (see fig. 6.1). While both species have discontinuous distributions, partly as a result of post-European land clearing, *E. sideroxylon*'s distribution is believed to have been more discontinuous pre-colonisation (Costermans, 1983). Despite their largely sympatric distributions, there appears to be some niche specialisation between these species, with *E. albens* occupying more fertile soils, and *E. sideroxylon* preferring drier, well-drained, more gravelly soils (Boland et al., 2006; Costermans, 1983; Harden, 2000). Despite their classification into different series, there is evidence of ongoing gene flow between these species, with reports of hybrid zones (Pryor, 1953), as is common in *Eucalyptus* generally, and especially in section Adnataria (Griffin et al., 1988).

### 6.3.2 Data acquisition

Samples used in this study were collected from naturally occurring trees of the target species throughout southeastern Australia. Leaf tissue and fruit were collected from between 3 and 15 trees from each location, across 39 distinct locations (fig. 6.1). Sample identifiers, GPS locations, and additional metadata are presented online (https://doi.org/10.6084/m9.figshare.7583291.v1). Sampling was performed between 2015 and 2017, primarily by Dr Jasmine Janes. Samples of species other than *E. albens* and *E. sideroxylon* that appear in Figure 6.3 are part of the same larger study, and will be desribed in a separate paper to be published imminently. Leaves were dried on silica gel, and 20-30 3 mm leaf hole punches were taken for DNA extraction (Harris Uni-Core

WB100039). Hole punches were added to 1.1 mL mini-tubes (Axygen Scientific) with a 3mm ball bearing, frozen under liquid nitrogen, and ground for 2 min using a TissueLyser (Qiagen). DNA extraction was performed using a 96-well column based kit, Invisorb DNA Plant HTS 96 Kit/ C 96 well purifications (Stratec Molecular 7037300400). The protocol was performed following the manufacturer's instructions, except for the lysis incubation, which was extended from 1 hour to 2 hours.

Multiplexed, short-read, whole-genome shotgun DNA sequencing libraries were generated using a cost-optimised, transposase-based protocol (Jones et al., 2018). Briefly, fluorometric DNA quantification was performed using a Quant-iT$^{TM}$ high sensitivity dsDNA assay kit (Molecular Probes$^{TM}$ Q33120). DNA was diluted to 2 ng/$\mu$L, quantified again and then diluted to 0.8 ng/$\mu$L, normalising concentrations across all samples. Then, 3 $\mu$L of each sample (2.24 ng) was transferred to a new plate with a small quantity of a Nextera$^{TM}$ tagment DNA enzyme (Illumina catalogue #15027865) to add adapters (tagmentation). This reaction was optimised to be 1/25th of manufacturer's protocol, to save reagents and increase throughput. Custom index primers were used to amplify the libraries during 13 cycles of PCR (primer sequences provided in Jones et al., 2018). Libraries were purified and size-selected with a combination of bead- and electrophoresis-based methods, selecting fragments with insert sizes between 200 and 500 bp. These purified libraries were sequenced on a variety of Illumina platforms, with most libraries sequenced on multiple runs across both NextSeq 500 and NovoSeq 2000 instruments at the Biomolecular Resource Facility, ANU, and the Ramaciotti Center, UNSW. Multiple runs were pooled by sample to obtain sufficient coverage. Library preparation was performed primarily by Dr Ashley Jones and Dr Norman Warthmann.

### 6.3.3 Alignment and polymorphism detection

Sequencing yielded between 3 Gbp and 10 Gbp per sample, pooled across all sequencing runs (see fig. 6.2). Raw sequence data was quality filtered using AdapterRemoval (Schubert et al., 2016), removing adaptor sequences, trimming low-quality ($<$Q25) sub-sequences, and merging overlapping read pairs. We used BWA MEM version 0.7.15 (Li, 2013; Li and Durbin, 2009) to align short reads using default alignment parameters to the *Eucalyptus grandis* reference genome (genome size 640Mbp), with an assembled *E. grandis* chloroplast added to the nuclear genome assembly (HM347959; Paiva et al., 2011). Across all samples, 90% percent of reads were aligned to the *E. grandis* reference, with an average alignment mismatch rate of 4.8%. Both read mapping and alignment mismatch rates suggest a reference bias between

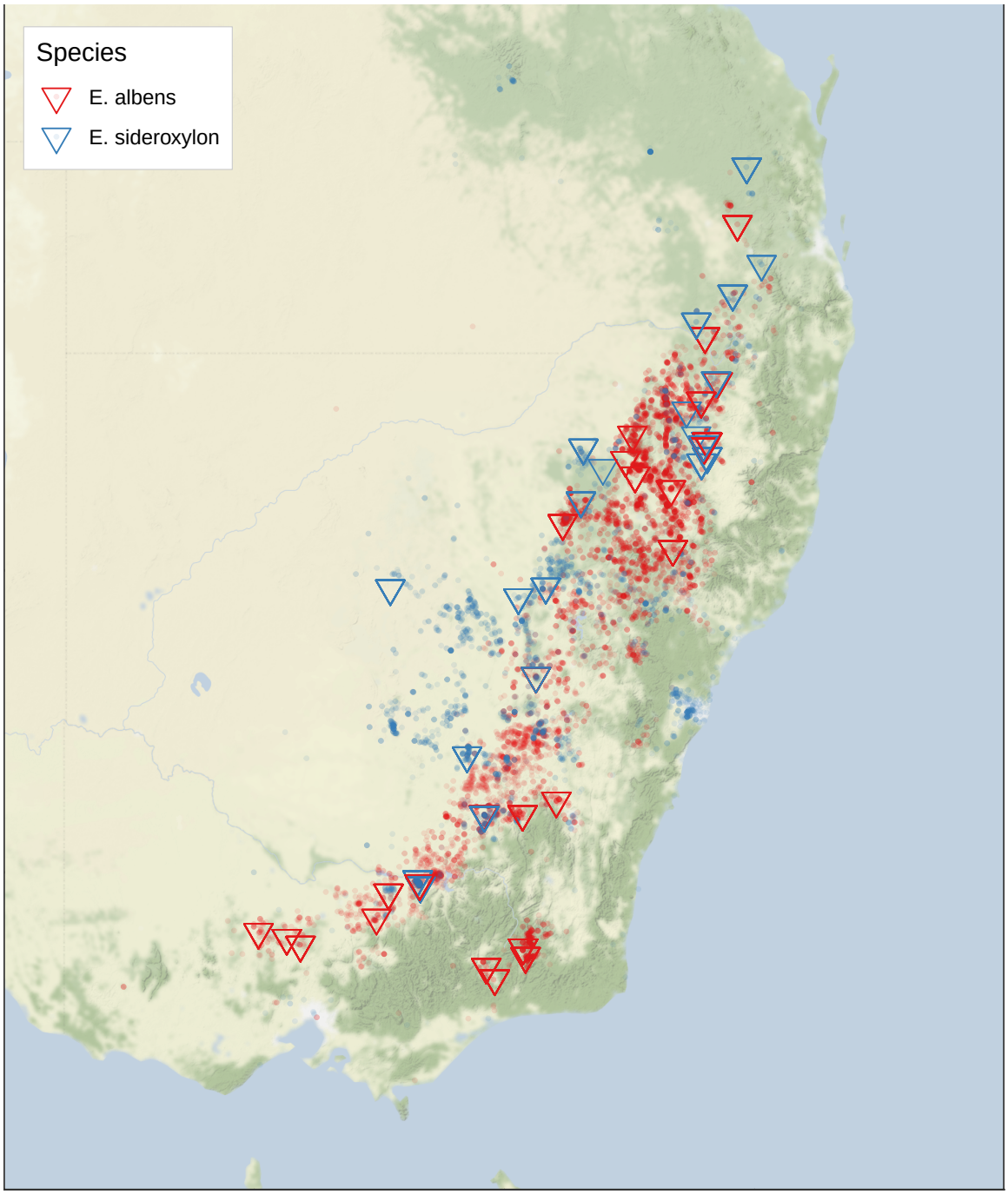



**Figure 6.1: Focal species occurrence records and sampling localities.** Geolocated occurrence records (±1 km accuracy) for *E. albens* and *E. sideroxylon* obtained from the Atlas of Living Australia are overlain on a map of southeastern Australia. Sampling localities used in this study are indicated by large triangles.

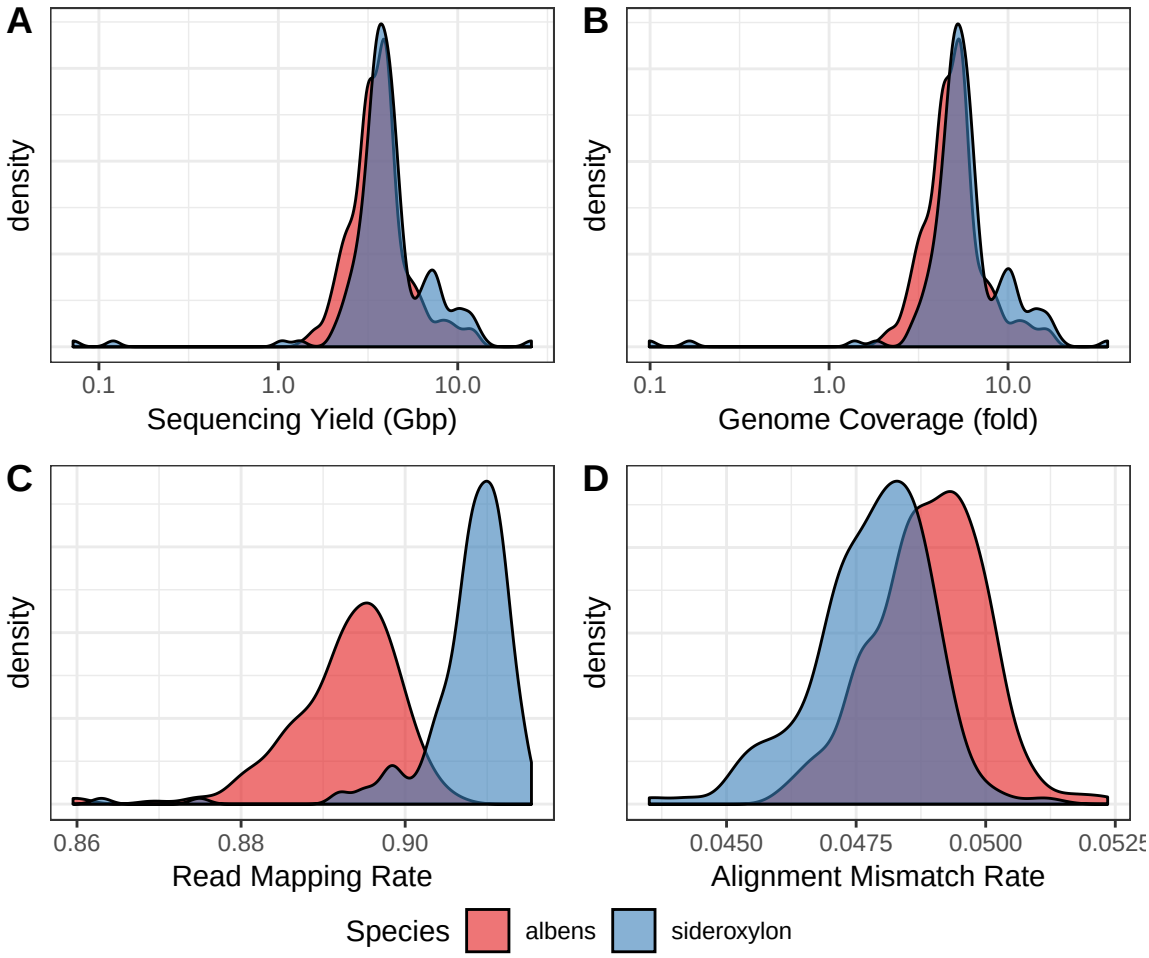



**Figure 6.2: Whole genome sequencing yield and alignment statistics.** A) Post-QC raw sequencing yield in bases, showing most samples yielded between 3 Gbp and 10 Gbp. B) *E. grandis* genome coverage (total sum of aligned bases). C) Read alignment rate, D), proportion of aligned bases which do not match the *E. grandis* reference genome. Overall, we have consistent, moderate coverage (median 5.2-fold), although both read mapping and alignment mismatch rates suggest a reference bias between species (with *E. sideroxylon* appearing less distant).

species (with *E. sideroxylon* appearing less distant).

We detected short genomic variants using an efficient pipeline implementing the variant calling models contained in FreeBayes (Garrison and Marth, 2012) and `bcftools mpileup` (Li, 2011). As these tools are not internally parallelised, and the volume of data generated in this project was very large, I developed a genomic region-parallelised system pipeline around these software. Briefly, this pipeline performs variant calling on each 100 kbp region of the *E. grandis* reference genome in parallel across hundreds of CPUs at once, before merging the candidate variants discovered in each region into a genome-wide variant set. This variant set was then normalised with `bcftools norm` (Li, 2011), block substitutions were decomposed to single nucleotide polymorphisms (SNPs) using `vt decompose_blocksub` (Tan et

al., 2015), and filtered with `bcftools filter`. We discarded variants with quality less than 10, fewer than five reads in total across all alleles in all samples, and fewer than three reads supporting the alternate allele across all samples. In total, we discovered 132 million putative variants, of which 55 million were common ($>10\%$ minor allele frequency) SNPs within at least one species.

While many analyses require knowledge of exact genotypes for each sample, some methods (e.g. ANGSD; Korneliussen et al., 2014) are able to represent uncertainty in individual genotypes through subsequent analyses. Given our low sequencing coverage, individual genotypes may have higher error than we desire, particularly in detecting heterozygosity. To address these concerns, we used ANGSD (Korneliussen et al., 2014) to detect putative variants, and to calculate genotype likelihoods at each variable site. ANGSD considered loci only if there were $>10$ reads at a SNP (summed across at least 10 samples with data), considered reads only if they had a mapping quality $>30$, considered bases within reads only if they had a base quality score $>20$, and removed variants with a minor allele frequency $<2\%$, with fewer than three reads supporting the alternate allele, or if the p-value of the likelihood-ratio test of non-zero minor allele frequency (i.e. test of polymorphism) was $>10^{-3}$. Indel and block-substitution variation is not considered by ANGSD. We used a region-parallel approach similar to that used in variant calling to accelerate this computation. In total, ANGSD detected 55 million polymorphisms (variants with $>=10\%$ minor allele frequency) across our samples.

From ANGSD likelihoods, we calculated several population genetic statistics. A two-dimensional site-frequency spectrum (SFS) between all *E. albens* and *E. sideroxylon* was calculated with realSFS (Nielsen et al., 2012), then estimated genome-wide $F_{ST}$ between *E. albens* and *E. sideroxylon* using this two-dimensional SFS as a prior (see Supplementary fig. 6.15). Using ngsDist (Fumagalli et al., 2014), we calculated inter-sample genetic distances for all samples that clustered into the two main species groups (based on kWIP distances). We estimated inter-sample covariance using PCAngsd (Meisner and Albrechtsen, 2018). We calculated Euclidean distances from PCAngsd covariances using the Gower transformation ($D_{ij} = C_{ii} + C_{jj} - 2C_{ij}$; Gower, 1985).

We implemented all steps in the above pipeline as a generic, modular workflow using the Snakemake workflow manager (Köster and Rahmann, 2012). Snakemake allows parallelisation of variant calling across genomic regions in a way that is abstracted from the execution environment. Project and cluster specific configuration of this pipeline is separate to pipeline code, allowing easy adaptation to other systems and datasets. In fact, this pipeline has sub-

sequently been used in at least three additional projects (wheat, tomato, and potato population genomics). This pipeline and associated scripts are open source, and available online at https://github.com/kdmurray91/euc-dp14-workspace.

### 6.3.4 Population genetic analysis

We performed kmer-based exploratory genetic analysis, to confirm sample identities and guide subsequent analyses. Genetic distances were estimated using kWIP, a kmer-based estimator of genetic distance described in Chapter 4 (and Murray et al., 2017). We first counted 21-mers in unaligned, quality trimmed sequencing reads, after pooling all reads for each sample into one file. We estimated inter-sample genetic distances using the weighted inner product metric implemented in kWIP, as it showed highest performance at low coverage (see Chapter 4). Distances were estimated on each data subset (all 10 Adnataria species, both *E. albens* and *E. sideroxylon*, and *E. albens* and *E. sideroxylon* separately) to allow subset-specific weighting. We visualised these exploratory analyses using both hierarchical clustering (`hclust`) and classical multidimensional scaling (`cmdscale`) in R 3.4 (R Core Team, 2018). In addition to kmer-based estimates of genetic distance, we visualised the sample covariance (or genomic relationship matrix) as estimated by PCAngsd in a similar fashion, and compared these results visually.

To examine within-locality diversity, a variety of population diversity metrics were employed. We calculated Nei's sample-size corrected gene diversity (or expected heterozygosity, $H_e = \frac{2N}{2N-1}\frac{1-p_l^2-q_l^2}{L}$; Nei and Roychoudhury, 1974), using per-locality allele frequencies calculated from expected genotypes by PCAngsd. Additionally, we calculated gene diversity for all pairs of sampling locations, by considering all individuals from both localities of a pair as a single site (equivalent to gene diversity in a sample-size weighted mean of allele frequencies). We displayed these measures of intra- and inter-location genetic diversity by plotting location estimates on a map of south-eastern Australia using ggmap and Stamen map layers (Kahle and Wickham, 2013).

Traditional model-based genetic clustering methods like STRUCTURE (Pritchard et al., 2000) and ADMIXTURE (Alexander et al., 2009) were designed to detect discrete population structure, therefore they may perform poorly for continuously distributed natural populations in which isolation by distance is the primary driver of genetic structure (Frantz et al., 2009). ConStruct addresses this limitation by jointly modeling the effects of both continuous isolation by distance and discrete population structure on inter-sample relationships (Bradburd et al., 2017). As we expected continuously distributed landscape features to con-

tribute to inter-sample genetic distances, we used conStruct to simultaneously test for discrete and continuous population structure. We used per-locality allele frequencies calculated from PCAngsd expected genotypes. We tested two distinct models separately for *E. albens* and *E. sideroxylon*, using the cross-validation approach implemented in conStruct: a model similar to that used by STRUCTURE, and one allowing for isolation by distance within genetic clusters ('layers'). Layer contributions were calculated for all cross-validation runs. To test for recent admixture between *E. sideroxylon* and *E. albens*, we used conStruct directly on the estimated genotypes, again performing cross-validation and calculating layer contributions.

We estimated the distribution of genome-wide linkage disequilibrium by calculating inter-SNP correlations and modeling correlation decay as a function of chromosomal position. Using the BoringLD R package (https://github.com/kdmurray91/boringld), we first calculated pairwise $r^2$ among SNPs in 30 kbp genomic windows with an overlap of 10 kbp between adjacent windows from FreeBayes-called variants. Then, we fitted analytical models of the decay of $r^2$ as a function of inter-SNP base pair distance to determine the recombination rate ($\rho$), using formulae derived by Hill and Weir (1988). The base pair distance to half-maximal $r^2$ was also calculated for each window. Window estimates of both $\rho$ and half-maximal $r^2$ were summarised across all genome windows.

### 6.3.5 Landscape genomic analyses

We used Generalised Dissimilarity Modelling (GDM) to test for isolation by distance without assuming a linear relationship between geographic and genetic distance using the `gdm` R package (Manion et al., 2018). Using genetic distances derived from PCAngsd covariance, we modeled genetic distance as a function of geographic distance within each species. We calculated geographic distances between samples with earth.dist from the `fossil` R package (Vavrek, 2011). Models were constructed using individual-level genetic and geographic distances, using three I-spline knots. Only distance pairs with a geographic distance greater than 10 kilometers (i.e. inter-location pairs) were considered. For each model, we examined the robustness of spline fits using jackknifing with 100 replicates. For each jackknife replicate, we removed all samples from a random 10% of sampling locations and fitted the GDM models as before. To perform cross-validation of each model, we partitioned data into training and test sets comprising 90% and 10% of sampling locations, respectively. We then computed cross-validation accuracy as the correlation between actual genetic distances for all distances for the 10% test data partition, and the corresponding distances predicted using a GDM model trained on samples from the 90% training data partition.

To assess isolation by environment, we first selected potentially relevant environmental variables based on a general methodology described by Williams et al. (2012). Variable values were extracted using the Atlas of Living Australia's (ALA) Spatial Portal ("Atlas of Living Australia," 2018). To determine which variables to include in models of IBE, we first performed forward selection within each category: Water, Energy, and Soil (see Supplementary tbl. 6.2). We excluded terrain and geoscientific variables, as these processes vary over finer spatial scales than our aggregated sampling resolution. In each forward selection run, we started with a GDM model of genetic distance as a function of geographic distance, and proceeded by adding the variable that, when included, increased the proportion of deviance explained by the model by the largest amount. We terminated this process when no variable could explain at least 1% of additional deviance. We then combined forward-selected variables across all categories into a candidate GDM model. To assess how representative our sampling was of each species' range, we compared distributions of each environmental variable from our sampling locations to distributions for ALA observation records for each species.

To refine candidate GDM models, and assess the importance and significance of constituent variables, we performed backward selection using the `gdm.VarImp` function in the GDM package (Ferrier et al., 2007; Manion et al., 2018), with 100 permutation replicates for each step. For both species, the inflection point in decreased model deviance explained resulted in five variables retained for the final model (Supplementary fig. 6.12). We then assessed the consistency of spline fits using the jack-knifing approach described above. These new functions for variable selection and cross-validation are available as an R package (https://github.com/kdmurray91/gdmhelpers).

## 6.4 Results

### 6.4.1 Population genetic variation

After filtering unsupported or singleton variants, we discovered over 100 million candidate variants (varying slightly between software tools; Supplementary tbl. 6.1). This equates to about 1/6th of all positions in the *E. grandis* reference genome. Of these candidate variants, around 40% were not segregating ($<10\%$ minor allele frequency) in either *E. albens* and *E. sideroxylon*. Of the remaining approximately 60 million variants, over half were segregating in both species, with 22% private to *E. albens* and 23% private to *E. sideroxylon* (Supplementary tbl. 6.1). ANGSD estimated inter-species genome-wide $F_{ST}$ between *E. albens* and *E. sideroxylon* to be 0.15; global intraspecific $F_{ST}$ was 0.018 in *E. albens* and 0.017 in *E. sideroxy-*

*lon*.

Using kmer-based estimators of genetic distance, we estimated genome-wide differentiation within the Adnataria. A principal component analysis (PCA) on kWIP distance estimates showed four clusters corresponding to taxonomic series. In some cases, species formed discrete subgroups within series, though in many cases species clusters overlapped somewhat (fig. 6.3); within-series divergence between species varies. Hierarchical clustering of kWIP distances showed similar patterns. The two focal species of this study formed clearly distinct clusters, as expected (Supplementary fig. 6.13).

*Eucalyptus albens* and *E. sideroxylon* had high genetic diversity. Expected heterozygosity within sampling locations ranged between 0.2 and 0.3 for both species, with *E. sideroxylon* having slightly lower mean location-level diversity, particularly in northern localities. Both species exhibited high species-wide genetic diversity (*E. sideroxylon* $H_e = 0.25$, $\pi = 0.053$; *E. albens* $H_e = 0.26$, $\pi = 0.056$). Background linkage disequilibrium (LD) decayed rapidly in both species (Supplementary fig. 6.11). The median base-pair distance to half-maximal $r^2$ in *E. albens* was 92 bp (IQR 47-219 bp), while LD extended slightly further in *E. sideroxylon* (median 113 bp; IQR 55-264 bp).

### 6.4.2 Spatial genetic diversity and structure

In general, genetic diversity was spread evenly over the range of our sampling in both species (fig. 6.4). Both $\pi$ and $H_e$ are almost equal across all locations sampled in *E. albens*, while genetic diversity in *E. sideroxylon* declined very slightly in locations toward the north of our sampling. Similarly, expected heterozygosity and $\pi$ among samples at pairs of locations were uncorrelated with pairwise geographic distance (Supplementary fig. 6.14).

### 6.4.3 No discrete but continuous population structure

Neither *E. albens* or *E. sideroxylon* exhibited strong signs of discrete population structure in a PCA of intra-sample genetic covariance as estimated by PCAngsd (fig. 6.5). Leading principal component axes explained little of the overall genomic variance between samples (0.8% and 0.6% in *E. albens*, 3.6% and 1.0% in *E. sideroxylon*). In each species, the leading principal component axis was correlated with latitude, suggesting isolation by geographic distance (*E. albens* $r^2 = 0.92$, $p < 0.0001$; *E. sideroxylon* $r^2 = 0.87$, $p < 0.0001$).

Joint estimation of continuous isolation by distance and discrete population structure indicated both species likely form single, continuous populations, with clinal structure influenced by strong IBD. When accounting for IBD in conStruct, cross-validation of conStruct

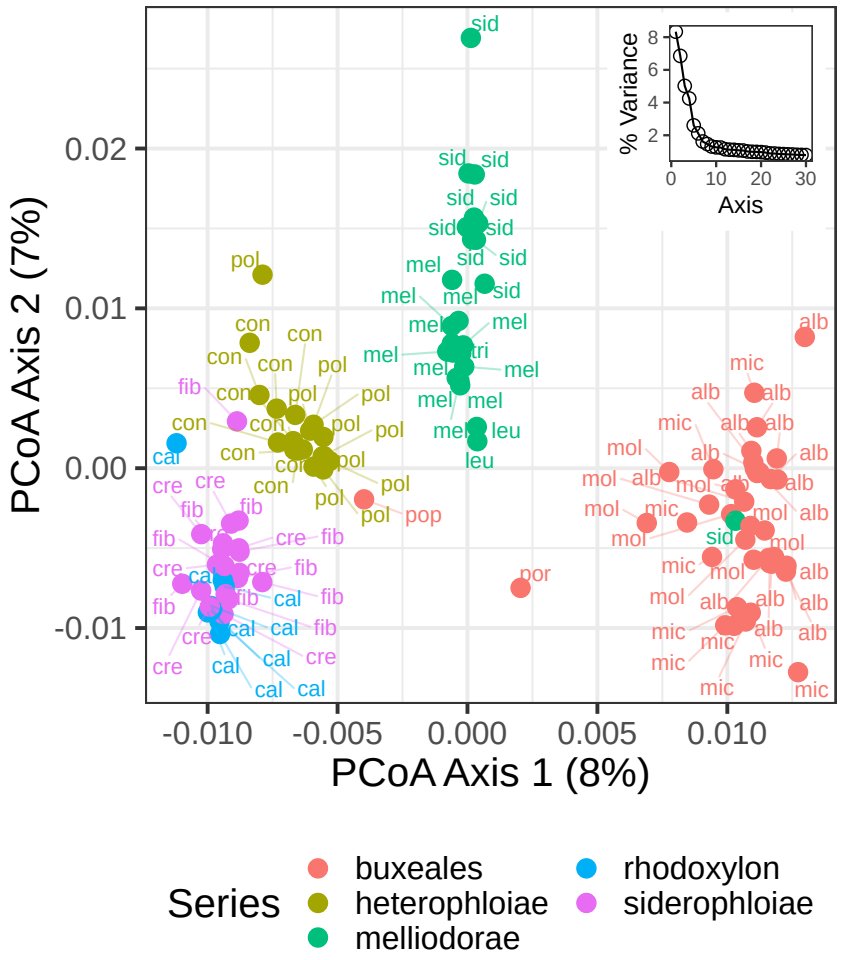



**Figure 6.3: Intra- and inter-series genomic divergence.** Principal coordinates analysis was performed on distances calculated using kWIP directly on short reads without alignment to a reference. Broadly, all samples from across five series in Adnataria form four clusters, corresponding to series-level divergences. Two series form a single cluster, Rhodoxlyon, and Siderophloiae; recent taxonomies reclassify *E. caleyi* within Siderophloiae (Nicolle, 2018). Within-series divergence between species varies. Individuals' species are denoted using the first three letters of species names: alb - *E. albens*, cal - *E. caleyi*, con - *E. conica*, cre - *E. crebra*, fib - *E. fibrosa*, leu - *E. leucoxylon*, mel - *E. melliodora*, mic - *E. microcarpa*, mol - *E. mollucana*, pol - *E. polyanthemos*, pop - *E. populnea*, sid - *E. sideroxylon*, tri - *E. tricarpa*. Samples appearing far from their species/series clusters likely represent either misidentification or sample mislabels, and were excluded from subsequent analyses.

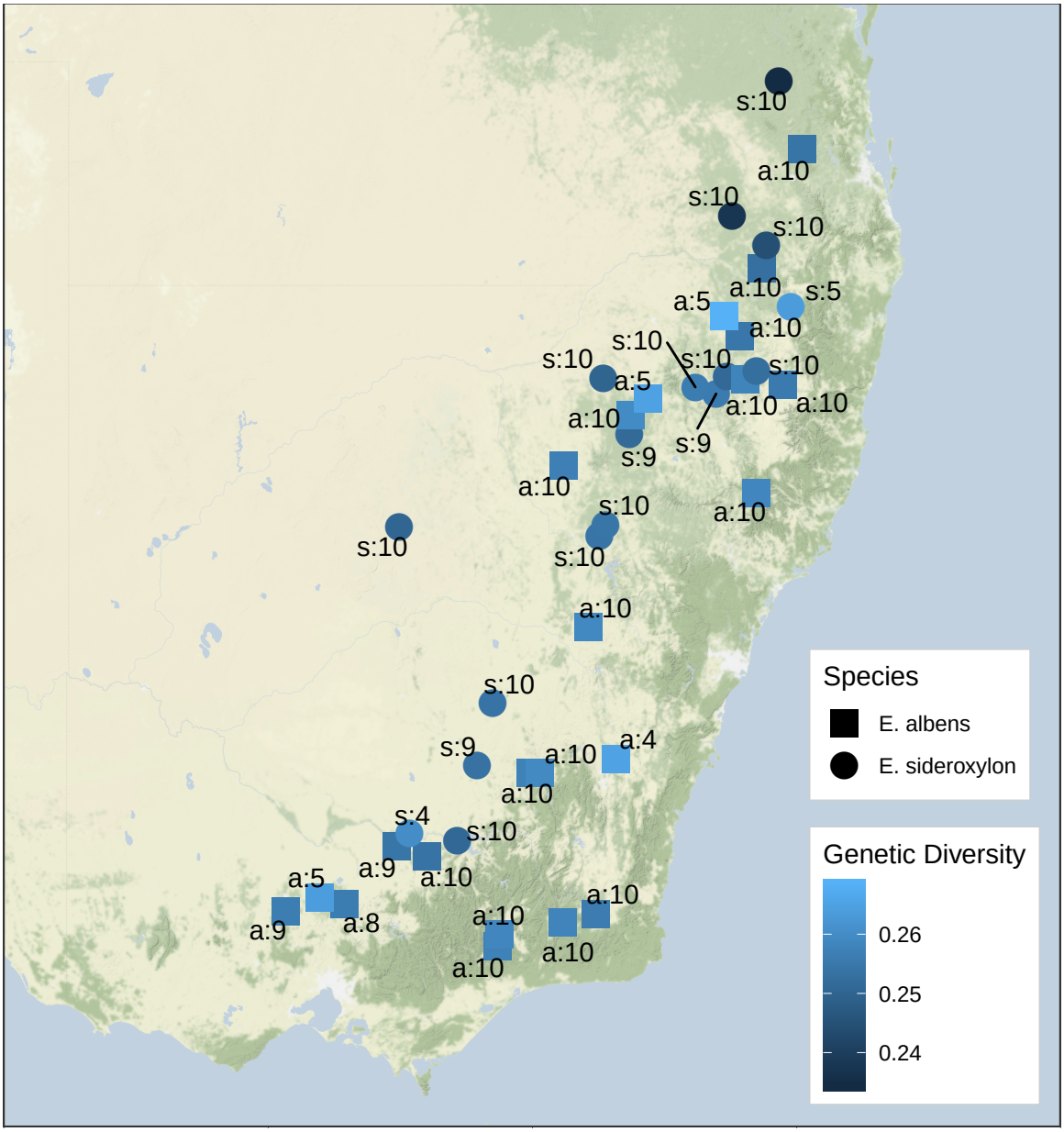



**Figure 6.4: Geographic surface of genetic diversity superimposed on a map of southeastern Australia.** Annotations describe species (s: or a: for *E. sideroxylon* and *E. albens* respectively) and the number of individuals per locality.

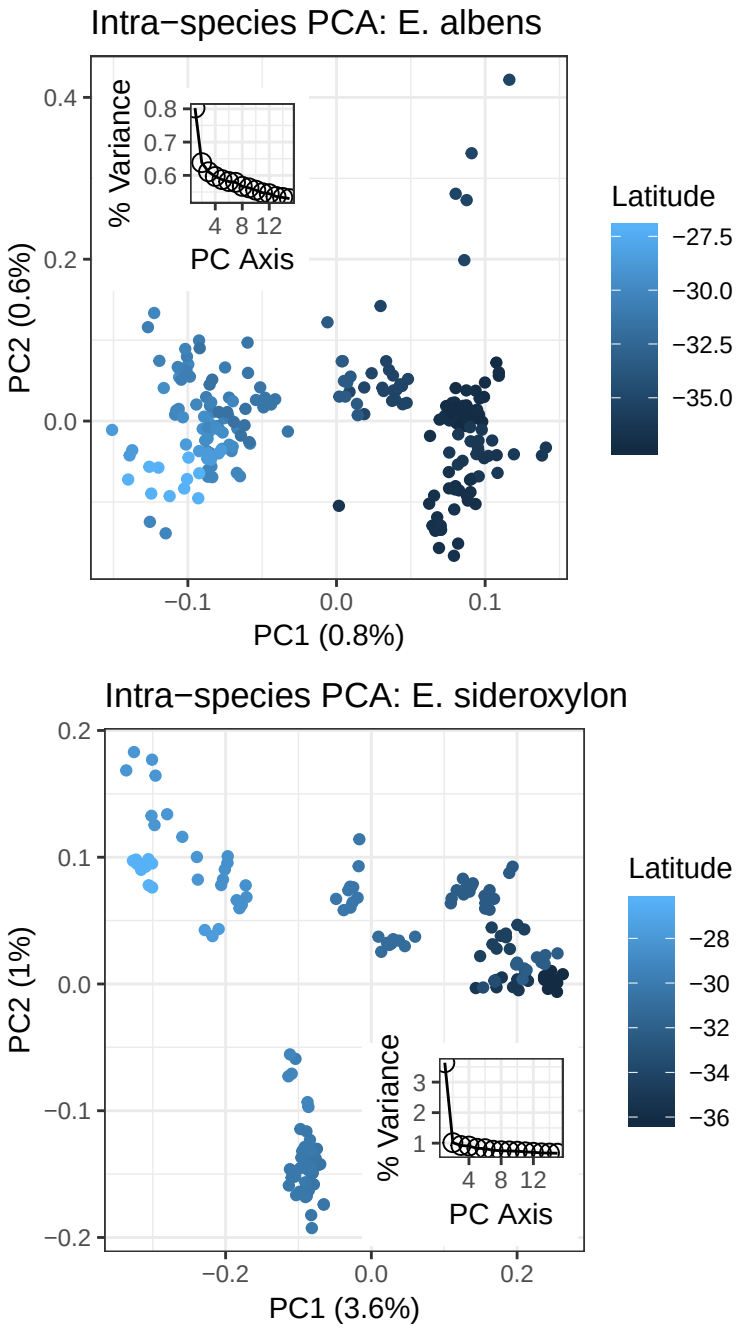



**Figure 6.5: Principal component analysis (PCA) of *E. albens* and *E. sideroxylon* individual genotypes.** Axes describe eigendecomposition of PCAngsd estimates of sample covariance. Individuals are coloured by latitude, the primary axis of variation in species' distributions. Insets show the distribution of leading eigenvalues. Note the absence of strong discrete clusters, the strong trend in PC1 across latitude, and the low proportion of genetic variance explained by each leading axis.

models suggested either one or two populations in both species (fig. 6.6). In models with two population layers, the second layer contributed very little additional predictive accuracy. The second layer in such models had no strong signal of IBD. This second layer could describe a small contribution of inter-species introgression to extant genetic diversity, or could represent "homogeneous minimum layer membership", an artifact produced by conStruct when there are significant levels of missing data (Bradburd et al., 2017). ConStruct models that did not allow continuous isolation by distance required at least two populations to achieve similar predictive accuracy (fig. 6.6).

#### Interspecific gene flow

We detected signals suggesting ongoing inter-species gene flow. Six samples were intermediate between *E. albens* and *E. sideroxylon*, being both intermediate in PCA (Supplementary fig. 6.13), and having interspecies admixture proportions between 30% and 70% (Supplementary fig. 6.16). Two of these samples were identified as putative hybrids in the field. Mantel tests of inter-species distance pairs showed weak but statistically significant correlation between genetic distance and geographic distance, indicating that co-located *E. albens* and *E. sideroxylon* had lower genetic distance than geographically distant samples. This pattern could be due to inter-series gene flow, and is not predicted by incomplete lineage sorting, but could also be caused by certain demographic histories (e.g., expansion from shared ancestral refugia). Individual admixture proportions estimated by conStruct models supported the status of these six samples as recent hybrids (Supplementary fig. 6.16). Additionally, conStruct models suggested a variable, small proportion (between 0% and 10%; Supplementary fig. 6.16) of admixture from *E. albens* to *E. sideroxylon* (or vice versa). Additionally, more than half of all variants that were common in either species were common in both species (Supplementary tbl. 6.1). These results concur with ABBA-BABA-based formal tests of admixture conducted in an as-yet-unpublished sister study (J. Janes, pers. comm.).

### 6.4.4 Isolation by distance and environment

Isolation by distance was moderately strong and largely linear in both species. Using generalised dissimilarity modelling (GDM) to model genetic distance as a function of geographic distance, we found *E. albens* to have moderately strong, almost linear IBD, with models explaining approximately 26% of overall deviance ($P < 0.001$; fig. 6.7). Meanwhile, *E. sideroxylon* exhibited very strong IBD, with models explaining 78% of overall deviance ($P < 0.001$; fig. 6.7). The relationships described by the best fit splines were robust to the removal of 10%

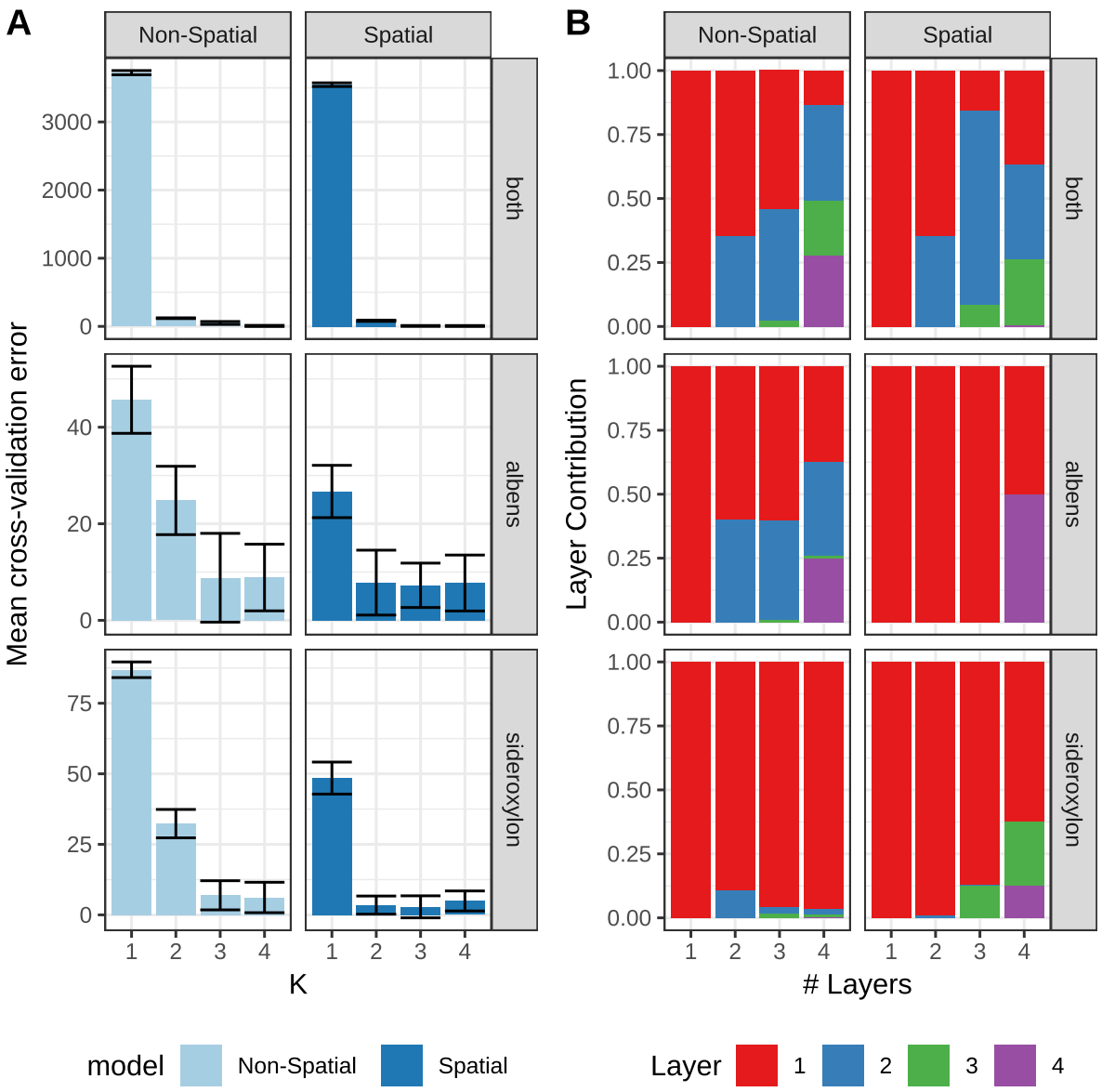



**Figure 6.6: Cross-validation of conStruct models of continuous and discrete population structure.** A) Model cross-validation error, means ± SD. B) Layer contribution to model explanatory power within each model with "# Layers". Non-spatial: construct models that do not account for IBD, Spatial: construct models that allow for IBD within each population layer. Plots rows are for datasets with all localities across both *E. albens* and *E. sideroxylon* ("both") or within each species.

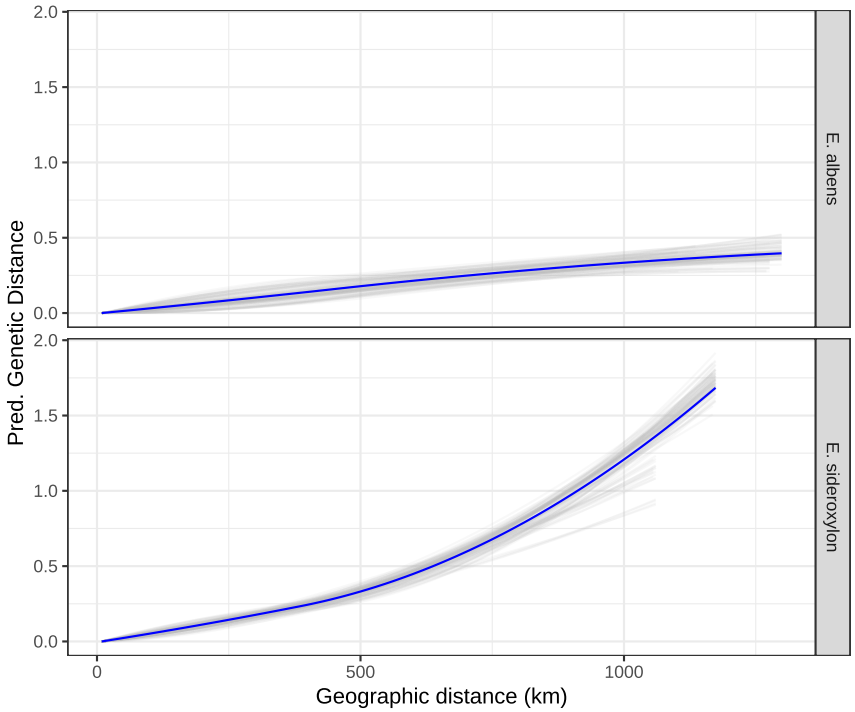



**Figure 6.7: Geographic GDM show strong isolation by distance.** GDM model splines (blue) and jackknife replicate splines (grey) that best describe the association between geographic distance and genetic distance in each species. Geography-only GDM models explain 26% of model deviance in *E. albens*, and 78% in *E. sideroxylon*. IBD appears to have an approximately linear trend in *E. albens*, while the strength of IBD increases for *E. sideroxylon* localities separated by more than 500 km.

of the sampling locations (i.e. jackknifing; fig. 6.7).

In the GDM analysis with environmental predictors, *E. albens* showed moderate isolation by environment, particularly driven by precipitation and substrate related environmental variables. Forward selection identified 11 candidate environmental covariates, each able to explain at least 1% additional deviance. Backward selection on these 11 variables identified substrate hydrological conductivity, substrate phosphorus concentration, spring/autumn precipitation seasonality, precipitation of the wettest quarter, and total wind run as contributing the highest predictive power (Supplementary tbl. 6.3). Overall, this model explained 31% of total deviance ($P < 0.001$), 7% higher than a model containing only geographic distance. Cross-validation showed this model to have reasonable predictive accuracy; the correlation between predicted and true genetic distances was $r^2 = 0.33$, roughly equal to the percentage of deviance explained (Supplementary fig. 6.10). For most variables, splines of best fit were robust to removal of 10% of sampling locations, although some variables had high uncertainty (e.g. precipitation of the wettest month), and other variables showed bimodal distributions of spline fits (e.g. autumn/spring precipitation seasonality; fig. 6.8).

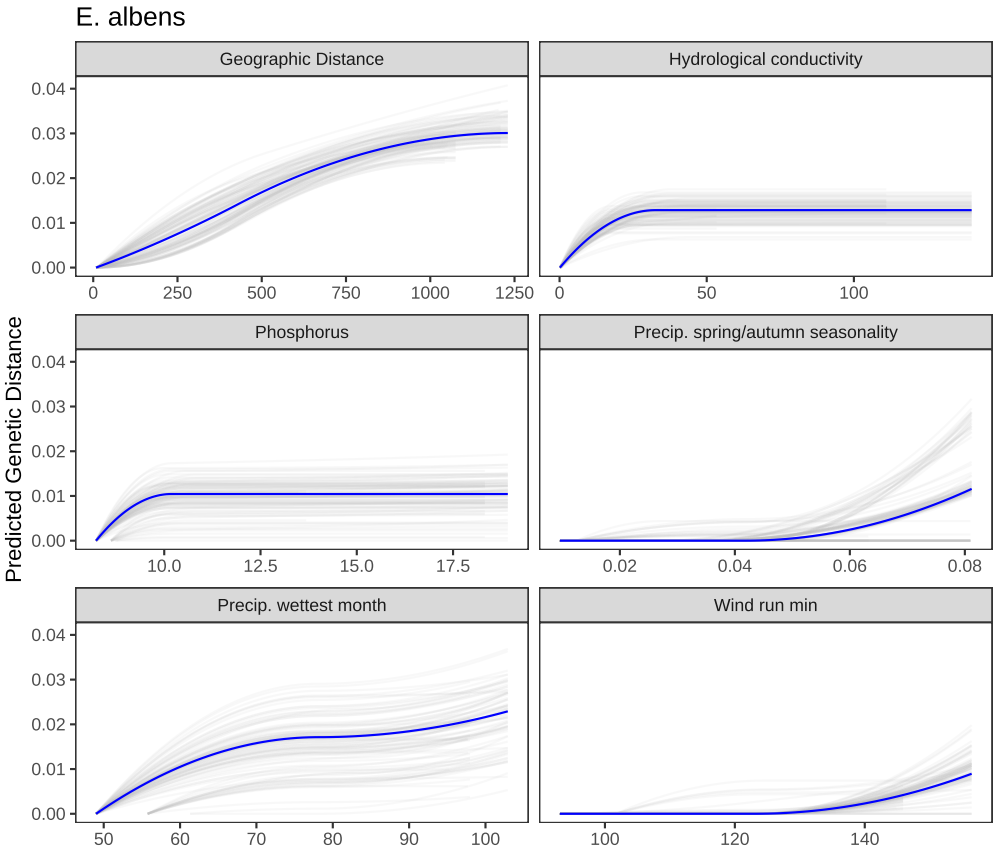



**Figure 6.8: GDM spline fits for *E. albens*.** To test the robustness of GDM predictive splines, models were re-run with 10% of sampling locations removed in each dataset. Each panel showed the range of spline fits among the 100 jackknife replicates (grey) and the full data (blue).

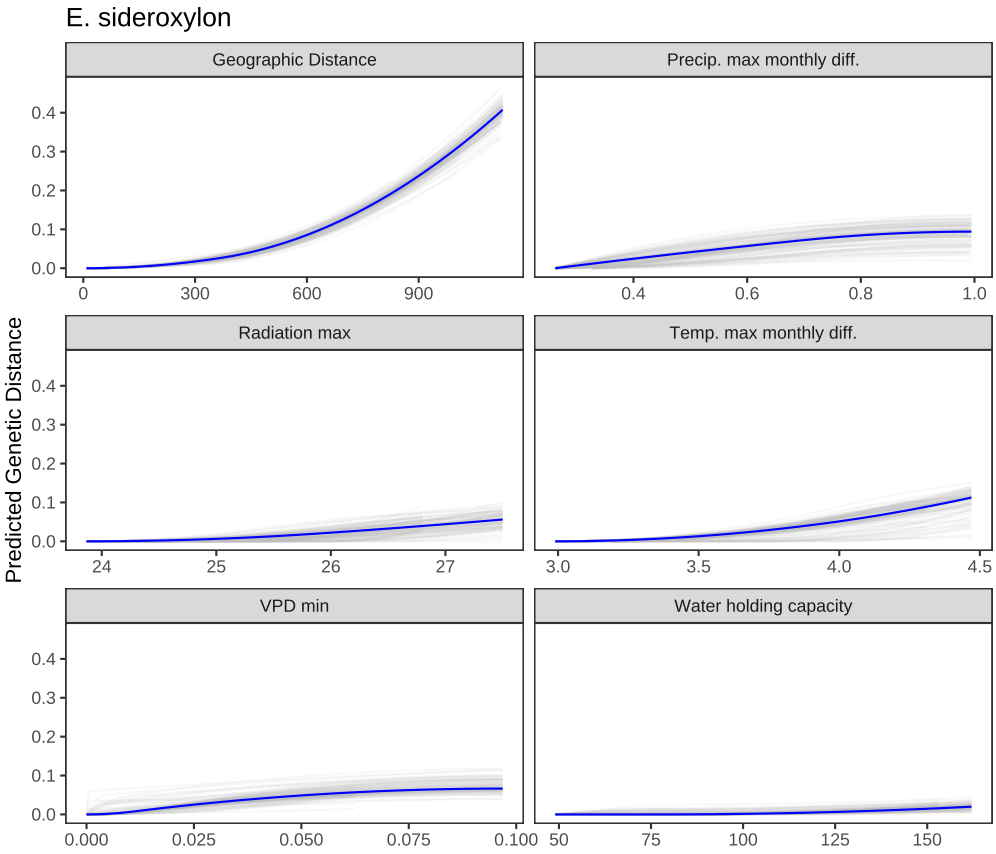



**Figure 6.9: GDM spline fits for *E. sideroxylon*.** To test the robustness of GDM predictive splines, models were re-run with 10% of sampling locations removed in each dataset. Each panel showed the range of spline fits among the 100 jackknife replicates (grey) and the full data (blue).

Similarly, *E. sideroxylon* showed somewhat stronger isolation by environment than *E. albens*, primarily driven by environmental variables describing the timing, availability, and demand for moisture. Forward selection identified 12 candidate covariates, and backward selection identified maximum cloud-adjusted solar radiation, maximum month-on-month differences in temperature and precipitation, maximal vapour pressure deficit, and substrate water holding capacity as the five variables with highest predictive power (Supplementary tbl. 6.3). Again, the overall model was highly significant ($P < 0.001$), explained 90% of total deviance (12% higher than a model containing only geographic distance), and had very high mean cross-validation predictive accuracy ($r^2 = 0.90$; Supplementary fig. 6.10). Splines of best fit were robust to removal of 10% of sampling locations for all predictors, with low uncertainty in spline fits across jack-knifing replicates fig. 6.9.

## 6.5 Discussion

### 6.5.1 Genetic diversity

Common, widespread eucalypts generally exhibit large, continuous populations with high genetic diversity and low population divergence. We confirm this result with one of the first whole-genome population resequencing studies in wild eucalypts (Kainer et al., 2018; Silva-Junior and Grattapaglia, 2015). We estimated intra-species $F_{ST}$ to be 0.017-0.018, lower than estimates from previous studies in a variety of eucalypt species (*E. melliodora*: $F_{ST} = 0.04$, Supple et al., 2018; *E. globulus*: $F_{ST} = 0.08$, Jones et al., 2002); although similar to estimates in other eucalypt species (*E. obliqua*: $F_{ST} = 0.015$, Bloomfield et al., 2011). These previous estimates are of similar magnitude to widespread tree species in other biomes, for example Oaks, Poplar, and Pine (*Quercus robur*: $F_{ST} = 0.07$, Vakkari et al., 2006; *Q. engelmannii*: $F_{ST} = 0.04$, Ortego et al., 2012; *Populus tremuloides*: $F_{ST} = 0.03$, Wyman et al., 2003; *Pinus taeda*: $F_{ST} = 0.04$, Eckert et al., 2010; *P. contorta*: $F_{ST} = 0.02$, Yang et al., 1996). This very weak genetic structure likely results from a combination of very large, stable effective population sizes, widespread ranges, and high outcrossing rates (Williams and Woinarski, 1997).

While high compared to many tree species, genetic diversity both across all individuals and within localities is slightly lower in *E. sideroxylon* than *E. albens*. Previous work indicated especially high allozyme diversity in *E. albens* (Prober and Brown, 1994). Estimates of effective population size within *E. albens* and *E. sideroxylon* follow a similar pattern (J. Janes *et al.*, in prep.). Linkage disequilibrium reported here is less extensive than in some previous reports (Silva-Junior and Grattapaglia, 2015), and is more similar to older estimates of LD decay from wild individuals of *E. grandis* (Grattapaglia and Kirst, 2008) and *E. globulus* (Thavamanikumar et al., 2011).

A crucial caveat to these results is that we predominantly sampled from mature trees which likely predate the extensive land clearing and habitat fragmentation that accompanied European colonisation of Australia. The applicability of these results and conclusions to future generations of these species is uncertain. Individuals from later generations show reduced but still high genetic and/or phenotypic diversity in recent studies of related *Eucalyptus* species (Broadhurst, 2013; Jordan et al., 2016; Supple et al., 2018), although these studies examined planted individuals, either in provenance trials or revegetation efforts (Costa e Silva et al., 2011). Further research on the differences in genetic diversity between remnant stands and younger cohorts is warranted.

### 6.5.2 Continuous genetic divergence

We observed continuous differentiation across the landscape within both species, driven both by geography and environment. This matches findings in most previous studies of genomic variation in eucalypts (Jordan et al., 2017; Steane et al., 2015, 2014; Supple et al., 2018). However, unlike previous studies, we found no support for strong discrete genetic structure. As seen in simulated and empirical studies of continuously distributed species (Bradburd et al., 2017; Frantz et al., 2009), we found statistical support for discrete population structure only when IBD was not incorporated into models of population structure. This conflation of IBD and discrete structure cements the conclusion that accurate determination of population structure in widespread species should use methods that can jointly estimate isolation by distance and discrete population structure.

We found very strong isolation by distance, particularly in *E. sideroxylon*. This is much stronger than in previous studies on related species at similar spatial scales. For example, weak isolation by distance occurs among populations in *E. melliodora*, with little correlation of genetic and geographic distance between pairs separated by less than 500 km (Supple et al., 2018; but see Andrew et al., 2005), and relatively weak IBD has been found in *E. microcarpa* (Jordan et al., 2017). Weak IBD may have technical and/or biological causes. Noisy reduced-representation sequencing methods that have large error in estimating sample genotypes (e.g. in *E. melliodora*; Supple et al., 2018), and therefore genetic distances, may have led to underestimation of the correlation between genetic and geographic distances. The difference in resolution in the present study may be partly due to our use of PCAngsd to calculate genetic distances, as it is designed to reduce the stochastic effects of low-coverage sequencing on inter-individual distances. Shirk et al. (2017) find distances based on PCA axes most accurately detect isolation by distance and environment, and PCAngsd is analogous to PCA-based distances in this context.

Strong IBD is likely a result of patterns of migration imposed by the reproductive ecology of eucalypts (Williams and Woinarski, 1997). Seed dispersal is limited in eucalypts, with pollen exchange accounting for the vast majority of migration among localities (Booth, 2017; Potts and Gore, 1995; Williams and Woinarski, 1997); recent analysis of chloroplast markers in box-ironbark eucalypts supports this (Alwadani et al., 2019). Pollination is facilitated by generalist insect, bird, and mammal pollinators in nearly all species (Potts and Gore, 1995; Williams and Woinarski, 1997). Most exchanges of pollen occur within a limited local range; however, migration events occur over much longer ranges with lower frequency (Williams and Woinarski, 1997). As a result, genes are readily exchanged far beyond immediate neigh-

bours. We found the strength of IBD to be strikingly different between *E. sideroxylon* and *E. albens*. This finding suggests that, while pollen-mediated gene flow is strong enough to limit discrete population structure in both species, gene flow at larger spatial scales is more restricted in *E. sideroxylon* than in *E. albens*. This goes against the expectation that the larger, more coloured flowers of *E. sideroxylon* attract more frequent bird pollination, leading to higher pollen motility. These observations are also supported by lower local genetic diversity within *E. sideroxylon*, particularly in northern localities.

### 6.5.3 Isolation by Environment

We observed isolation by environment in both species, primarily driven by variables describing the availability of water and nutrients to plants, with little influence of temperature. Permutation-based variable testing showed only a small orthogonal contribution of environment to observed genetic distances, after accounting for geographic distance. Strong spatial autocorrelation of environment variables prevents fully disentangling geographic and environmental contributions to gene flow across the landscape. Exclusion of relevant environmental variables could cause underestimation of overall IBE, although the variable selection procedure employed here tested the contribution of a broad range of environmental variables concerning soil, geology, precipitation, temperature, wind, solar radiation, and aridity. In most cases, inference of the environmental drivers of genomic differentiation appear robust to subsampling of localities. GDM models of isolation by distance and environment had high cross-validation accuracy, and all were significant under locality-wise permutation testing. While specific environmental variables selected as most important were not shared, the strength of IBE was similar in both species. Furthermore, the variables most predictive of genetic distance in both species described the availability and demand for moisture or soil fertility (nutrient or water availability). Despite local niche separation (Boland et al., 2006; Brooker and Kleinig, 2006; Costermans, 1983; Harden, 2000), the ranges of *E. albens* and *E. sideroxylon* overlap significantly (fig. 6.1), and therefore likely experience selection along similar macro-scale clines (e.g. temperature, aridity).

Correlation of genetic and environmental variation is well established in Eucalyptus. Differences in climate and soil nitrogen can predict genetic differentiation in *E. melliodora* (Supple et al., 2018). Allele frequencies at certain SNPs were significantly correlated with aridity, temperature, and rainfall in *E. tricarpa* (Steane et al., 2014), *E. loxophleba* (Steane et al., 2017a), and *E. microcarpa* (Jordan et al., 2017). Our use of environmental variables designed to interrogate the ecology of Australian plants (Williams et al., 2012) precludes direct comparison

of IBE among studies at the level of specific variables. However, our results follow a similar general pattern to these previous studies of gene-environment association in eucalypts.

### 6.5.4 Interspecific Divergence and Gene Flow

About half of all common variants discovered in this study are common in both species, and we observed low genome-wide divergence between *E. albens* and *E. sideroxylon* ($F_{ST} = 0.15$). Recent evidence suggests the genetic divergence is not strong at most genomic loci in many species, both in eucalypts (Rutherford et al., 2018) and more broadly (Andrew and Rieseberg, 2013; Wu, 2001). Additionally, low interspecific differentiation is expected theoretically given extremely large effective population sizes, long generation times, and relatively recent radiation (González-Orozco et al., 2016).

Interspecific gene flow between eucalypts has been observed many times, though probably occurs at a low rate in nature (Griffin et al., 1988). We made several observations suggestive of ongoing gene flow between *E. albens* and *E. sideroxylon* (Supplementary fig. 6.13; fig. 6.16). We identified several putative hybrid individuals in the field, via PCA, and conStruct indicated a low but consistent proportion of inter-series admixture. Hybridisation between *E. albens* and *E. sideroxylon* has been demonstrated previously (Pryor, 1953), and more broadly, a systematic review by Griffin et al. (1988) showed species within Eucalyptus section Adnataria were found to hybridise at the highest rate of any section. The proportion of hybrids we observe here is of the same approximate magnitude as that observed in several other eucalypts in the subgenus *Symphomyrtus* (1-3%; Williams and Woinarski, 1997). Hybridisation between *E. albens* and *E. sideroxylon* occurs in spite of ecological differentiation, for example, in the form of limited local co-occurrence, different tolerance of poor soils and aridity (Boland et al., 2006; Costermans, 1983; Harden, 2000), and relatively little overlap in flowering period (*E. albens*: January-June, *E. sideroxylon* May-November; Costermans, 1983; Brooker and Kleinig, 2006).

### 6.5.5 Conservation implications

To avoid extirpation, organisms must either adapt or migrate as environments change (Aitken et al., 2008). Our findings of high genetic diversity imply a large pool of variation accessible to natural selection. However, the long generation time of these trees makes it unlikely that natural selection on local standing variation alone can outpace anthropogenic changes in climate and land use; therefore, migration of better-adapted alleles is required (Booth, 2017; Booth et al., 2015). While we show pollen must have been exchanged over relatively large

distances at a rate historically sufficient to prevent strong differentiation between localities, natural rates of migration are unlikely to be prevent range contractions (Booth, 2017; Prober et al., 2015). Human assistance may be required to shift the ranges of these and many other woodland species (Butt et al., 2013; González-Orozco et al., 2016; Supple et al., 2018).

Management interventions can take numerous forms. There is a temptation to use models of isolation by environment to guide selection of seed sources for assisted migration. However, we urge the utmost caution when doing so: these models of IBE are based on genome-wide patterns among predominantly near-neutral genetic variation, and use predicted, interpolated environmental data. Such models could detect the historical influence of environment on genetic diversity, but there is no promise that these influences reflect what may happen in the future. In particular, we strongly discourage the use of these results (or the results of any similar study) to narrow the range of seed sources used to revegetate any given locality. Studies of inbreeding in eucalypts find strong effects of selfing and local inbreeding (Hardner and Potts, 1995), but little outbreeding depression was observed beyond hundreds of meters among intraspecific crosses of (Hardner et al., 1998). Outbreeding depression is observed in more distant crosses (e.g. by Lopez et al., 2000; Larcombe et al., 2016). Such results reinforce the need for a restoration strategy that focuses on adaptive potential as much as pre-adapted germplasm. Our advice matches that proposed in numerous recent syntheses of revegetation strategy (Broadhurst et al., 2008; Kardos and Shafer, 2018; Prober et al., 2015; Weeks et al., 2011), in particular "climate-adjusted provenancing" (Prober et al., 2015). As an additional consideration, climate change is not the only anthropogenic risk to these species: the habitat these species inhabit has been cleared extensively since European colonisation of Australia, with only a few percent of the habitat remaining (NSW Scientific Committee, 2002). Perhaps the most effective management action would be the prevention of further deforestation and habitat fragmentation, both for these species and generally.

### 6.5.6 Future directions

All patterns reported here concern genome-wide average effects; significant variation between loci in patterns described here likely exists. Investigating how variation in ancestry, population structure, interspecific differentiation, and associations with environment differ across the genome requires whole-genome datasets, and the dataset and analysis pipeline we present here enables these analyses. In particular, our finding of low linkage disequilibrium implies that many reduced-representation sequencing methods would provide data for just a fraction of all independent loci, and therefore miss important segregating variation (Ahrens et al.,

2018; Lowry et al., 2017).

Genotype-environment association (GEA) studies could detect individual alleles which vary in frequency across some environmental cline, accounting for geography and genome-wide patterns (as has been observed with reduced representation sequencing in related species, e.g. Steane et al., 2014, 2017a, 2017b). Loci that have undergone selective sweeps could also be detected, shedding further light on recent evolution (Nielsen et al., 2005). Similarly, investigation of inter-species divergence at specific loci could highlight which loci are maintaining species boundaries in the face of gene flow (Strasburg et al., 2012). Finally, genome-wide average ancestry may differ significantly from local ancestry at nearly all loci across the genome, and could be examined in these species (e.g. using Local PCA; Li and Ralph, 2018).

### 6.5.7 Conclusions

In summary, we found high intraspecific genetic diversity, low genome-wide divergence between *E. albens* and *E. sideroxylon*, and evidence of ongoing gene flow between these species. We found no evidence of strong, discrete population structure, and uncovered strong continuous isolation by distance in both species. We also found that isolation by geographic distance accounts for most, but not all, of this continuous genetic structure, with environmental variables describing the availability and demand for moisture, temperature, and substrate contributing to the pattern of IBE. Taken together, these results describe *E. albens* and *E. sideroxylon* as widespread species with high genetic diversity and strong isolation by distance. A small proportion of genetic variation is associated with climate; however, high levels of genetic diversity exist regionally, and even within localities. This high genetic diversity implies these species have high adaptive potential, especially if enhanced by assisted migration. The crucial test of these species' survival will not be the level of understanding we gain about the intricacies of isolation by landscape, but rather the extent to which we utilise these and other species in large-scale rehabilitation of degraded ecosystems.

#### Acknowledgements

We thank Norman Warthmann, Tim Collins, Jamieson Gorrell, Jeremy Bruhl, and Allison Huesler for technical assistance. This work was supported financially by the Australian Research Council (CE140100008; DP150103591), and an Australian Government Research Training Program scholarship. The research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

## 6.6 References


Ahrens CW, Rymer PD, Stow A, Bragg J, Dillon S, Umbers KDL, Dudaniec RY. **2018**. "The search for loci under selection: Trends, biases and progress." *Molecular Ecology* **27**:1342–1356. doi:10.1111/mec.14549

Aitken SN, Yeaman S, Holliday JA, Wang T, CurtisMcLane S. **2008**. "Adaptation, migration or extirpation: Climate change outcomes for tree populations." *Evolutionary Applications* **1**:95–111. doi:10.1111/j.1752-4571.2007.00013.x

Alexander DH, Novembre J, Lange K. **2009**. "Fast model-based estimation of ancestry in unrelated individuals." *Genome Res* **19**:1655–1664. doi:10.1101/gr.094052.109

Alwadani KG, Janes JK, Andrew RL. **2019**. "Chloroplast genome analysis of box-ironbark Eucalyptus." *Molecular Phylogenetics and Evolution* **136**:76–86. doi:10.1016/j.ympev.2019.04.001

Andrew RL, Peakall R, Wallis IR, Foley WJ. **2007**. "Spatial Distribution of Defense Chemicals and Markers and the Maintenance of Chemical Variation." *Ecology* **88**:716–728. doi:10.1890/05-1858

Andrew RL, Peakall R, Wallis IR, Wood JT, Knight EJ, Foley WJ. **2005**. "Marker-Based Quantitative Genetics in the Wild?: The Heritability and Genetic Correlation of Chemical Defenses in Eucalyptus." *Genetics* **171**:1989–1998. doi:10.1534/genetics.105.042952

Andrew RL, Rieseberg LH. **2013**. "Divergence Is Focused on Few Genomic Regions Early in Speciation: Incipient Speciation of Sunflower Ecotypes." *Evolution* **67**:2468–2482. doi:10.1111/evo.12106

Atlas of Living Australia. **2018**. `https://www.ala.org.au/`

Bloomfield JA, Nevill P, Potts BM, Vaillancourt RE, Steane DA. **2011**. "Molecular genetic variation in a widespread forest tree species Eucalyptus obliqua (Myrtaceae) on the island of Tasmania." *Aust J Bot* **59**:226–237. doi:10.1071/BT10315

Boland DJ, Brooker MIH, Chippendale GM, Hall N, Hyland BPM, Johnston RD, Kleinig DA, McDonald MW, Turner JD. **2006**. "Forest Trees of Australia." Victoria, Australia: Csiro Publishing.

Booth TH. **2017**. "Going nowhere fast: A review of seed dispersal in eucalypts." *Aust J Bot* **65**:401–410. doi:10.1071/BT17019

Booth TH et al. **2015**. "Native forests and climate change: Lessons from eucalypts." *Forest Ecology and Management* **347**:18–29. doi:10.1016/j.foreco.2015.03.002

Bradburd G, Coop G, Ralph P. **2017**. "Inferring Continuous and Discrete Population Genetic Structure Across Space." *bioRxiv* 189688. doi:10.1101/189688

Broadhurst LM. **2013**. "A genetic analysis of scattered Yellow Box trees (Eucalyptus mel-

liodora A.Cunn. Ex Schauer, Myrtaceae) and their restored cohorts." *Biological Conservation* **161**:48–57. doi:10.1016/j.biocon.2013.02.016

Broadhurst LM, Lowe A, Coates DJ, Cunningham SA, McDonald M, Vesk PA, Yates C. **2008**. "Seed supply for broadscale restoration: Maximizing evolutionary potential." *Evolutionary Applications* **1**:587–597. doi:10.1111/j.1752-4571.2008.00045.x

Brooker I, Kleinig D. **2006**. "Field guide to eucalypts." Melbourne: Bloomings Books.

Butt N, Pollock LJ, McAlpine CA. **2013**. "Eucalypts face increasing climate stress." *Ecology and Evolution* **3**:5011–5022. doi:10.1002/ece3.873

Costa e Silva J, Hardner C, Tilyard P, Potts BM. **2011**. "The effects of age and environment on the expression of inbreeding depression in Eucalyptus globulus." *Heredity* **107**:50–60. doi:10.1038/hdy.2010.154

Costermans L. **1983**. "Native trees and shrubs of south-eastern Australia." Sydney: Reed.

Dillon S, McEvoy R, Baldwin DS, Rees GN, Parsons Y, Southerton S. **2014**. "Characterisation of Adaptive Genetic Diversity in Environmentally Contrasted Populations of Eucalyptus camaldulensis Dehnh. (River Red Gum)." *PLOS ONE* **9**:e103515. doi:10.1371/journal.pone.0103515

Eckert AJ, Bower AD, GonzálezMartínez SC, Wegrzyn JL, Coop G, Neale DB. **2010**. "Back to nature: Ecological genomics of loblolly pine (Pinus taeda, Pinaceae)." *Molecular Ecology* **19**:3789–3805. doi:10.1111/j.1365-294X.2010.04698.x

Ferrier S, Drielsma M, Manion G, Watson G. **2002**. "Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling." *Biodiversity and Conservation* **11**:2309–2338. doi:10.1023/A:1021374009951

Ferrier S, Manion G, Elith J, Richardson K. **2007**. "Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment." *Diversity and Distributions* **13**:252–264. doi:10.1111/j.1472-4642.2007.00341.x

Frantz AC, Cellina S, Krier A, Schley L, Burke T. **2009**. "Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: Clusters or isolation by distance?" *Journal of Applied Ecology* **46**:493–505. doi:10.1111/j.1365-2664.2008.01606.x

Fumagalli M, Vieira FG, Linderoth T, Nielsen R. **2014**. "ngsTools: Methods for population genetics analyses from next-generation sequencing data." *Bioinformatics* **30**:1486–1487. doi:10.1093/bioinformatics/btu041

Garrison E, Marth G. **2012**. "Haplotype-based variant detection from short-read sequencing."

González-Orozco CE et al. **2016**. "Phylogenetic approaches reveal biodiversity threats under climate change." *Nature Climate Change* **6**:1110–1114. doi:10.1038/nclimate3126

Gower JC. **1985**. "Properties of Euclidean and non-Euclidean distance matrices." *Linear Algebra and its Applications* **67**:81–97. doi:10.1016/0024-3795(85)90187-9

Grattapaglia D, Kirst M. **2008**. "Eucalyptus applied genomics: From gene sequences to breeding tools." *New Phytologist* **179**:911–929. doi:10.1111/j.1469-8137.2008.02503.x

Griffin AR, Burgess IP, Wolf L. **1988**. "Patterns of Natural and Manipulated Hybridisation in the Genus Eucalyptus L'hérit. - A Review." *Aust J Bot* **36**:41–66. doi:10.1071/bt9880041

Guillot G, Rousset F. **2013**. "Dismantling the Mantel tests." *Methods in Ecology and Evolution* **4**:336–344. doi:10.1111/2041-210x.12018

Harden GJ. **2000**. "Flora of New South Wales." UNSW Press.

Hardner CM, Potts BM. **1995**. "Inbreeding depression and changes in variation after selfing in Eucalyptus globulus ssp. Globulus." *Silvae Genetica* **44**:46–54.

Hardner CM, Potts BM, Gore PL. **1998**. "The relationship between cross success and spatial proximity of Eucalyptus globulus ssp. Globulus parents." *Evolution* **52**:614–618. doi:10.1111/j.1558-5646.1998.tb01660.x

Hill WG, Weir BS. **1988**. "Variances and covariances of squared linkage disequilibria in finite populations." *Theoretical Population Biology* **33**:54–78. doi:10.1016/0040-5809(88)90004-4

Hoffmann A et al. **2015**. "A framework for incorporating evolutionary genomics into biodiversity conservation and management." *Climate Change Responses* **2**:1. doi:10.1186/s40665-014-0009-x

Jones A, Borevitz J, Warthmann N. **2018**. "Cost-conscious generation of multiplexed short-read DNA libraries for whole genome sequencing v1 (protocols.io.unbevan)." doi:10.17504/protocols.io.unbevan

Jones RC, Steane DA, Potts BM, Vaillancourt RE. **2002**. "Microsatellite and morphological analysis of Eucalyptus globulus populations." *Can J For Res* **32**:59–66. doi:10.1139/x01-172

Jones TH, Vaillancourt RE, Potts BM. **2007**. "Detection and visualization of spatial genetic structure in continuous Eucalyptus globulus forest." *Molecular Ecology* **16**:697–707. doi:10.1111/j.1365-294X.2006.03180.x

Jordan R, Dillon SK, Prober SM, Hoffmann AA. **2016**. "Landscape genomics reveals altered genome wide diversity within revegetated stands of Eucalyptus microcarpa (Grey

Box)." *New Phytol* **212**:992–1006. doi:10.1111/nph.14084

Jordan R, Hoffmann AA, Dillon SK, Prober SM. **2017**. "Evidence of genomic adaptation to climate in Eucalyptus microcarpa: Implications for adaptive potential to projected climate change." *Molecular Ecology* **26**:6002–6020. doi:10.1111/mec.14341

Kahle D, Wickham H. **2013**. "Ggmap: Spatial Visualization with ggplot2." *The R Journal* **5**:144–161.

Kainer D, Stone EA, Padovan A, Foley WJ, Külheim C. **2018**. "Accuracy of Genomic Prediction for Foliar Terpene Traits in Eucalyptus polybractea." *G3: Genes, Genomes, Genetics* **8**:2573–2583. doi:10.1534/g3.118.200443

Kardos M, Shafer ABA. **2018**. "The Peril of Gene-Targeted Conservation." *Trends in Ecology & Evolution* **33**:827–839. doi:10.1016/j.tree.2018.08.011

Korneliussen TS, Albrechtsen A, Nielsen R. **2014**. "ANGSD: Analysis of Next Generation Sequencing Data." *BMC Bioinformatics* **15**:356. doi:10.1186/s12859-014-0356-4

Köster J, Rahmann S. **2012**. "Snakemake — a scalable bioinformatics workflow engine." *Bioinformatics* **28**:2520–2522. doi:10.1093/bioinformatics/bts480

Larcombe MJ, Costa e Silva J, Tilyard P, Gore P, Potts BM. **2016**. "On the persistence of reproductive barriers in Eucalyptus: The bridging of mechanical barriers to zygote formation by F1 hybrids is counteracted by intrinsic post-zygotic incompatibilities." *Ann Bot* **118**:431–444. doi:10.1093/aob/mcw115

Li H. **2013**. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM."

Li H. **2011**. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data." *Bioinformatics* **27**:2987–2993. doi:10.1093/bioinformatics/btr509

Li H, Durbin R. **2009**. "Fast and accurate short read alignment with Burrows–Wheeler transform." *Bioinformatics* **25**:1754–1760. doi:10.1093/bioinformatics/btp324

Li H, Ralph PL. **2018**. "Local PCA Shows How the Effect of Population Structure Differs Along the Genome." *Genetics* genetics.301747.2018. doi:10.1534/genetics.118.301747

Lopez GA, Potts BM, Tilyard PA. **2000**. "F 1 hybrid inviability in Eucalyptus : The case of E. Ovata Œ E. Globulus." *Heredity* **85**:242. doi:10.1046/j.1365-2540.2000.00739.x

Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, Storfer A. **2017**. "Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation." *Mol Ecol Resour* **17**:142–152. doi:10.1111/1755-0998.12635

Manion G, Lisk M, Ferrier S, Nieto-Lugilde D, Mokany K, Fitzpatrick MC. **2018**. "Gdm:

Generalized Dissimilarity Modeling."

Meisner J, Albrechtsen A. **2018**. "Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data." *Genetics* **210**:719–731. doi:10.1534/genetics.118.301336

Murray KD, Webers C, Ong CS, Borevitz J, Warthmann N. **2017**. "kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity." *PLOS Computational Biology* **13**:e1005727. doi:10.1371/journal.pcbi.1005727

Nei M, Roychoudhury AK. **1974**. "Sampling variances of heterozygosity and genetic distance." *Genetics* **76**:379–390.

Nicolle D. **2018**. "Classification of the eucalypts (Angophora, Corymbia and Eucalyptus), Version 3." Adelaide, Australia: Dean Nicolle.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. **2012**. "SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data." *PLOS ONE* **7**:e37558. doi:10.1371/journal.pone.0037558

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. **2005**. "Genomic scans for selective sweeps using SNP data." *Genome Res* **15**:1566–1575. doi:10.1101/gr.4252305

NSW Scientific Committee. **2002**. "White box, yellow box, and Blakely's red gum woodland - endangered ecological community listing." Sydney: New South Wales Government.

Ortego J, Riordan EC, Gugger PF, Sork VL. **2012**. "Influence of environmental heterogeneity on genetic diversity and structure in an endemic southern Californian oak." *Molecular Ecology* **21**:3210–3223. doi:10.1111/j.1365-294X.2012.05591.x

Paiva JA et al. **2011**. "Advancing Eucalyptus genomics: Identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries." *BMC Genomics* **12**:137. doi:10.1186/1471-2164-12-137

Potts BM, Gore PL. **1995**. "Reproductive biology and controlled pollination of Eucalyptus-a review." University of Tasmania.

Pritchard JK, Stephens M, Donnelly P. **2000**. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* **155**:945–959.

Prober SM, Brown AHD. **1994**. "Conservation of the Grassy White Box Woodlands: Population Genetics and Fragmentation of Eucalyptus albens." *Conservation Biology* **8**:1003–1013. doi:10.1046/j.1523-1739.1994.08041003.x

Prober SM, Byrne M, McLean EH, Steane DA, Potts BM, Vaillancourt RE, Stock WD. **2015**. "Climate-adjusted provenancing: A strategy for climate-resilient ecological restoration." *Front Ecol Evol* **3**. doi:10.3389/fevo.2015.00065

Pryor LD. **1953**. "Anther shape in Eucalyptus genetics and systematics." *Proceedings of the Linnean Society of New South Wales* **78**:43–48.

Pryor LD, Johnson LAS. **1971**. "A classification of the Eucalypts." Canberra: Australian National University.

R Core Team. **2018**. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing.

Rutherford S, Rossetto M, Bragg JG, McPherson H, Benson D, Bonser SP, Wilson PG. **2018**. "Speciation in the presence of gene flow: Population genomics of closely related and diverging Eucalyptus species." *Heredity* **121**:126–141. doi:10.1038/s41437-018-0073-2

Schubert M, Lindgreen S, Orlando L. **2016**. "AdapterRemoval v2: Rapid adapter trimming, identification, and read merging." *BMC Research Notes* **9**:88. doi:10.1186/s13104-016-1900-2

Shirk AJ, Landguth EL, Cushman SA. **2017**. "A comparison of individual-based genetic distance metrics for landscape genetics." *Molecular Ecology Resources* **17**:1308–1317. doi:10.1111/1755-0998.12684

Silva-Junior OB, Grattapaglia D. **2015**. "Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis." *New Phytol* **208**:830–845. doi:10.1111/nph.13505

Spear SF, Balkenhol N, Fortin M-J, Mcrae BH, Scribner K. **2010**. "Use of resistance surfaces for landscape genetic studies: Considerations for parameterization and analysis." *Molecular Ecology* **19**:3576–3591. doi:10.1111/j.1365-294X.2010.04657.x

Steane DA, Conod N, Jones RC, Vaillancourt RE, Potts BM. **2006**. "A comparative analysis of population structure of a forest tree, Eucalyptus globulus (Myrtaceae), using microsatellite markers and quantitative traits." *Tree Genetics & Genomes* **2**:30–38. doi:10.1007/s11295-005-0028-7

Steane DA, Mclean EH, Potts BM, Prober SM, Stock WD, Stylianou VM, Vaillancourt RE, Byrne M. **2017a**. "Evidence for adaptation and acclimation in a widespread eucalypt of semi-arid Australia." *Biol J Linn Soc* **121**:484–500. doi:10.1093/biolinnean/blw051

Steane DA, Potts BM, McLean E, Collins L, Prober SM, Stock WD, Vaillancourt RE, Byrne M. **2015**. "Genome-wide scans reveal cryptic population structure in a dry-adapted eucalypt." *Tree Genetics & Genomes* **11**:33. doi:10.1007/s11295-015-0864-z

Steane DA, Potts BM, McLean EH, Collins L, Holland BR, Prober SM, Stock WD, Vaillancourt RE, Byrne M. **2017b**. "Genomic Scans across Three Eucalypts Suggest that

Adaptation to Aridity is a Genome-Wide Phenomenon." *Genome Biol Evol* **9**:253–265. doi:10.1093/gbe/evw290

Steane DA, Potts BM, McLean E, Prober SM, Stock WD, Vaillancourt RE, Byrne M. **2014**. "Genome-wide scans detect adaptation to aridity in a widespread forest tree species." *Mol Ecol* **23**:2500–2513. doi:10.1111/mec.12751

Strasburg JL, Sherman NA, Wright KM, Moyle LC, Willis JH, Rieseberg LH. **2012**. "What can patterns of differentiation across plant genomes tell us about adaptation and speciation?" *Philos Trans R Soc Lond B Biol Sci* **367**:364–373. doi:10.1098/rstb.2011.0199

Supple MA, Bragg JG, Broadhurst LM, Nicotra AB, Byrne M, Andrew RL, Widdup A, Aitken NC, Borevitz JO. **2018**. "Landscape genomic prediction for restoration of a Eucalyptus foundation species under climate change." *eLife* **7**:e31835. doi:10.7554/eLife.31835

Tan A, Abecasis GR, Kang HM. **2015**. "Unified representation of genetic variants." *Bioinformatics* **31**:2202–2204. doi:10.1093/bioinformatics/btv112

Thavamanikumar S, McManus LJ, Tibbits JFG, Bossinger G. **2011**. "The significance of single nucleotide polymorphisms (SNPs) in Eucalyptus globulus breeding programs." *Australian Forestry* **74**:23–29. doi:10.1080/00049158.2011.10676342

Thornhill AH, Ho SYW, Külheim C, Crisp MD. **2015**. "Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny." *Molecular Phylogenetics and Evolution* **93**:29–43. doi:10.1016/j.ympev.2015.07.007

Vakkari P, Blom A, Rusanen M, Raisio J, Toivonen H. **2006**. "Genetic variability of fragmented stands of pedunculate oak (Quercus robur) in Finland." *Genetica* **127**:231–241. doi:10.1007/s10709-005-4014-7

Vavrek MJ. **2011**. "Fossil: Palaeoecological and palaeogeographical analysis tools." *Palaeontologia Electronica* **14**:1T.

Wang IJ, Bradburd GS. **2014**. "Isolation by environment." *Mol Ecol* **23**:5649–5662. doi:10.1111/mec.12938

Weeks AR et al. **2011**. "Assessing the benefits and risks of translocations in changing environments: A genetic perspective." *Evol Appl* **4**:709–725. doi:10.1111/j.1752-4571.2011.00192.x

Williams JE, Woinarski J. **1997**. "Eucalypt ecology: Individuals to ecosystems." Cambridge; New York: Cambridge University Press.

Williams KJ, Belbin L, Austin MP, Stein JL, Ferrier S. **2012**. "Which environmental variables should I use in my biodiversity model?" *International Journal of Geographical Information Science* **26**:2009–2047. doi:10.1080/13658816.2012.698015

Wright S. **1943**. "Isolation by Distance." *Genetics* **28**:114–138.

Wu C-I. **2001**. "The genic view of the process of speciation." *Journal of Evolutionary Biology* **14**:851–865. doi:10.1046/j.1420-9101.2001.00335.x

Wyman J, Bruneau A, Tremblay MF. **2003**. "Microsatellite analysis of genetic diversity in four populations of Populus tremuloides in Quebec." *Can J Bot* **81**:360–367. doi:10.1139/b03-021

Yang R-C, Yeh FC, Yanchuk AD. **1996**. "A Comparison of Isozyme and Quantitative Genetic Variation in Pinus contorta ssp. Latifolia by FST." *Genetics* **142**:1045–1052.

Zeller KA, McGarigal K, Whiteley AR. **2012**. "Estimating landscape resistance to movement: A review." *Landscape Ecol* **27**:777–797. doi:10.1007/s10980-012-9737-0

## 6.7 Supplementary information

**Table 6.1:** SNP genotyping statistics.

| Segregating in | Angsd SNPs |
|---|---|
| Neither species | 78017065 |
| Both species | 29409038 |
| *E. albens* | 12407339 |
| *E. sideroxylon* | 12644030 |
| Total | 132477472 |

**Table 6.2: Environmental variables considered in forward selection of IBE models.**

| Abbrev. Name | Williams *et al.* Class | Name | Description |
|---|---|---|---|
| maxti | Energy | Temperature - coolest month max | Maximum temperature coolest month (řC) |
| maxtx | Energy | Temperature - month hottest maximum | Maximum temperature hottest month (řC) |
| minti | Energy | Temperature - coldest month min | Minimum temperature coldest month (řC) |
| mintx | Energy | Temperature - warmest month min | Minimum temperature warmest month (řC) |
| radni | Energy | Radiation - min month precipitation modified | Minimum month rainfall-modified solar radiation (MJ/m2/day) |
| radnx | Energy | Radiation - max month precipitation modified | Maximum month rainfall-modified solar radiation (MJ/m2/day) |
| rh2max | Energy | Humidity - month max relative | Maximum month relative humidity (%) |
| rh2min | Energy | Humidity - month min relative | Minimum month relative humidity (%) |
| rtimax | Energy | Temperature - max difference in min | Maximum difference in minimum temperatures (řC/day) |
| rtimin | Energy | Temperature - min difference in min | Minimum difference in minimum temperatures (řC/day) |
| rtxmax | Energy | Temperature - max difference in max | Maximum difference in maximum temperatures (řC/day) |
| rtxmin | Energy | Temperature - min difference in max | minimum difference in maximum temperatures (řC/day) |
| tmaxabsx | Energy | Temperature - max absolute mean max | Maximum month absolute mean maximum temperature (řC) |
| tminabsi | Energy | Temperature - min absolute mean min | Minimum month absolute mean minimum temperature (řC) |

| Abbrev. Name | Williams *et al.* Class | Name | Description |
|---|---|---|---|
| trngi | Energy | Temperature - min month diurnal range | Minimum month diurnal temperature range (řC) |
| trngx | Energy | Temperature - max month diurnal range | Maximum month diurnal temperature range (řC) |
| vpd2max | Energy | Vapour pressure deficit - month max | Maximum month vapour pressure deficit (KPa) |
| vpd2min | Energy | Vapour pressure deficit - month min | Minimum month vapour pressure deficit (KPa) |
| wind_windmi | Energy | Wind run - month min | Wind run - month min (km/day) |
| wind_windmx | Energy | Wind run - month max | Wind run - month max (km/day) |
| wind_windspmax | Energy | Wind speed - month max 9am or 3pm | Wind speed - month max 9am or 3pm (m/s) |
| wind_windspmin | Energy | Wind speed - month min 9am or 3pm | Wind speed - month min 9am or 3pm (m/s) |
| substrate_bdensity | Soil | Bulk density | Solum average bulk density (Mg/m3) |
| substrate_calcrete | Soil | Calcrete | Calcrete in or below soil profile (presence) |
| substrate_clay | Soil | Clay % | Solum average median clay content (%) |
| substrate_coarse | Soil | Soils - coarse | Soils dominated by coarse fragments including ironstone (class) |
| substrate_ksoilerr | Soil | Hydrological conductivity - uncertainty | Solum average uncertainty of horizon saturated hydraulic conductivity estimates (index) |
| substrate_ksat | Soil | Hydrologic conductivity - average saturated | Solum average median horizon saturated hydraulic conductivity (mm/h) |
| substrate_nmnlconcn0 | Soil | Nitrogen concentration pre-European | Pre-European estimate of mean annual concentration of mineral nitrogen in soil water (NMnlConc0.Base) |

| Abbrev. Name | Williams *et al.* Class | Name | Description |
|---|---|---|---|
| substrate_nstor0 | Soil | Nitrogen - plant-available pre-European | Pre-European estimate of mean annual store of total plant-available soil nitrogen (NTot0.Base) |
| substrate_nutrients | Soil | Nutrient status | Gross nutrient status (rating) |
| substrate_pmnlconcn0 | Soil | Phosphorus pre-European | Pre-European estimate of mean annual concentration of dissolved phosphorus in soil water (PMnlConc0.Base) |
| substrate_pmnln0 | Soil | Phosphorus - plant-available pre-European | Pre-European estimate of mean annual store of plant-available mineral phosphorus (PMnl0.Base) |
| substrate_soildepth | Soil | Soil depth | Solum depth (surface and subsoil layers) (metres) |
| substrate_soilpawhc | Soil | Water holding capacity - plant-available | Plant-available soil water holding capacity (mm) |
| substrate_wo_unr | Soil | Unreliable water retention parameters | Solum average unreliable water retention parameters (index) |
| adefi | Water | Precipitation deficit - month max | Maximum month precipitation deficit (mm) |
| adefx | Water | Precipitation deficit - month min | Minimum month precipitation deficit (mm) |
| arid_max | Water | Aridity index - month max | Maximum month aridity index |
| arid_min | Water | Aridity index - month min | Minimum month aridity index |
| evapi | Water | Evaporation - month min | Minimum month evaporation (mm) |
| evapx | Water | Evaporation - month max | Maximum month evaporation (mm) |
| raini | Water | Precipitation - driest month | Precipitation of the driest month (mm) |

| Abbrev. Name | Williams *et al.* Class | Name | Description |
|---|---|---|---|
| rainx | Water | Precipitation - wettest month | Precipitation of the wettest month (mm) |
| rprecmax | Water | Precipitation - max difference between successive months | Greatest rainfall difference between successive months (mm/day) |
| rprecmin | Water | Precipitation - min difference between successive months | Least rainfall difference between successive months (mm/day) |
| slrain0 | Water | Precipitation - annual (log) seasonality index | annual (log) rainfall seasonality index |
| slrain1 | Water | Precipitation - summer or winter (log) season | summer or winter (log) rainfall season |
| slrain2 | Water | Precipitation - spring or autumn (log) season | Spring or autumn (log) rainfall season |
| srain0mp | Water | Precipitation - annual seasonality ratio | annual rainfall seasonality ratio |
| srain1mp | Water | Precipitation - solstice seasonality ratio | Solstice rainfall seasonality ratio |
| srain2mp | Water | Precipitation - equinox seasonality ratio | Equinox rainfall seasonality ratio |

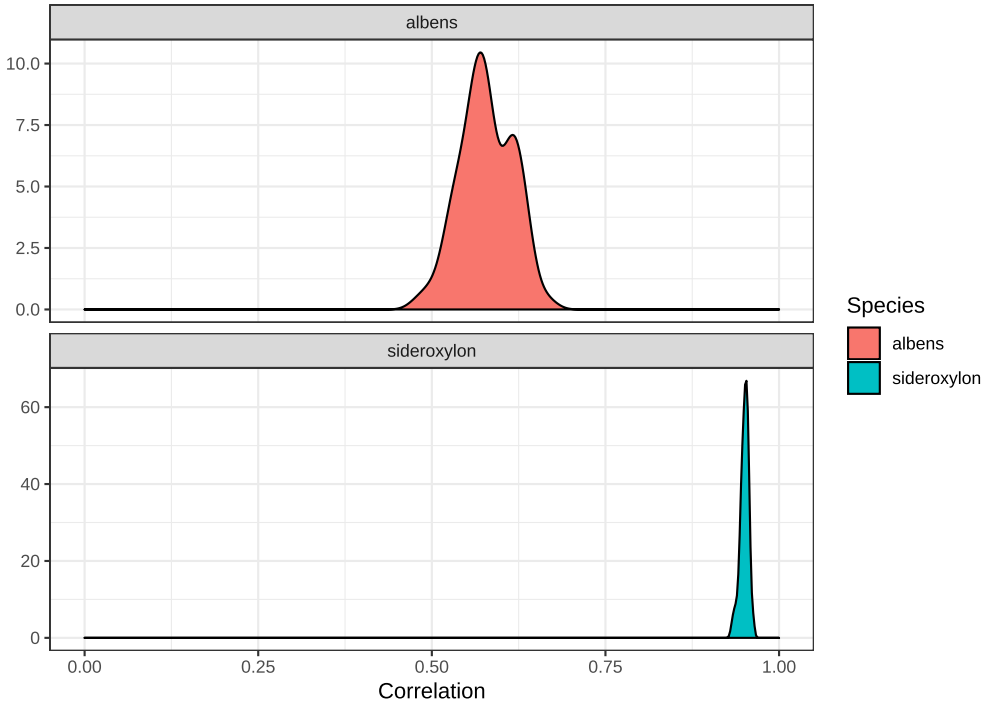



**Figure 6.10: Cross-validation accuracy of best-fit GDM models for *E. albens* and *E. sideroxylon*.** To test the predictive power of GDM models, GDM are fit on a training dataset with 10% of sampling locations removed in each dataset. The genetic distances of the remaining 10% of samples are predicted from their geographic and environmental data. Pearson's correlation is used to assess the goodness-of-fit between predicted and actual genetic distances.

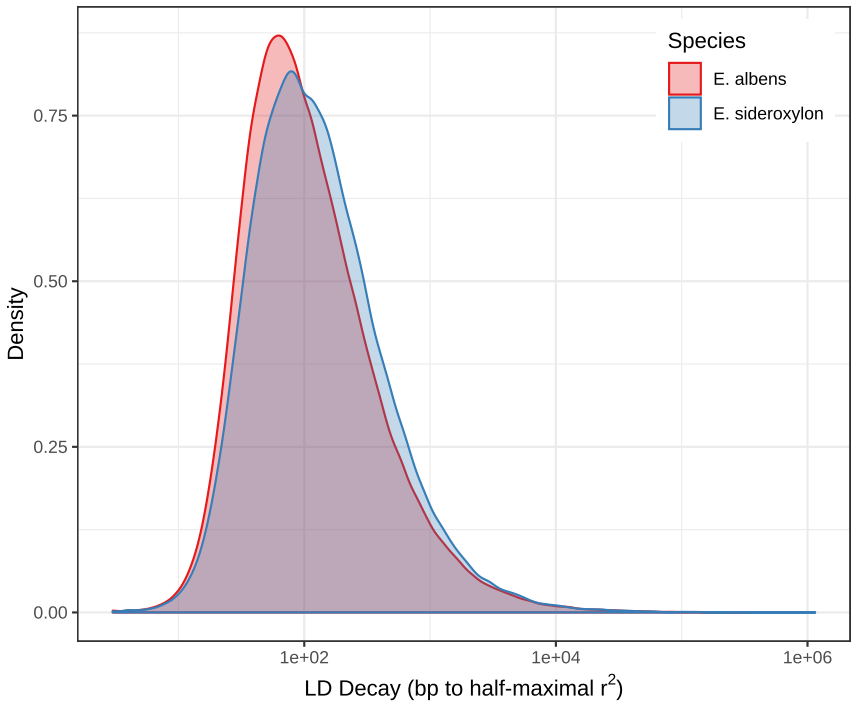



**Figure 6.11: Distribution of LD extent for *E. albens* and *E. sideroxylon*.** Here we show the distribution of LD extent, defined as the distance required for half-maximal decay in $R^2$, aggregated for all 1000000 bp genome windows

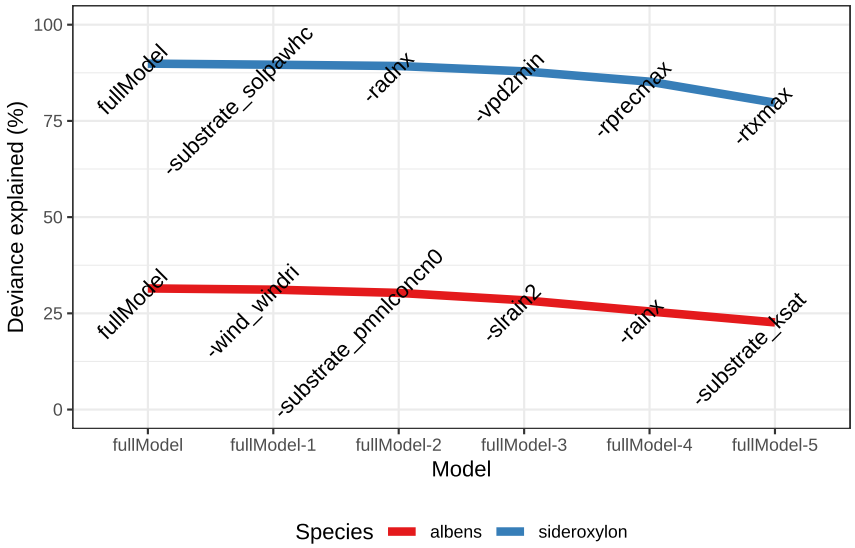



**Figure 6.12: GDM model deviance explained during back-selection of variables.** "fullModel" describes the model with all variables included. Each subsequent point removes one variable (per labels on plot). Please see supplementary tbl. 6.2 for variable names

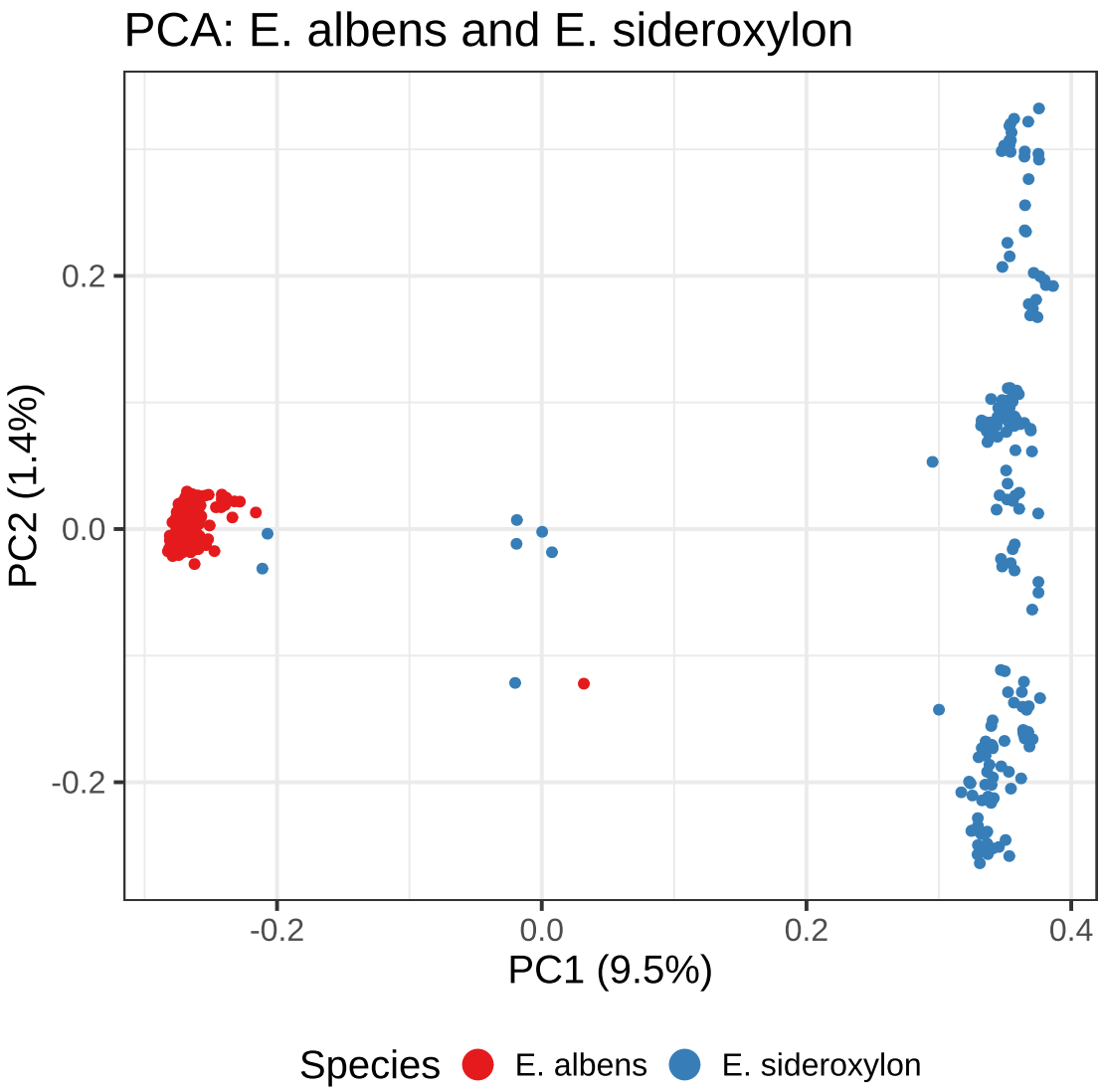



**Figure 6.13: Cross-species PCA of genetic covariance estimated from PCANGSD.**

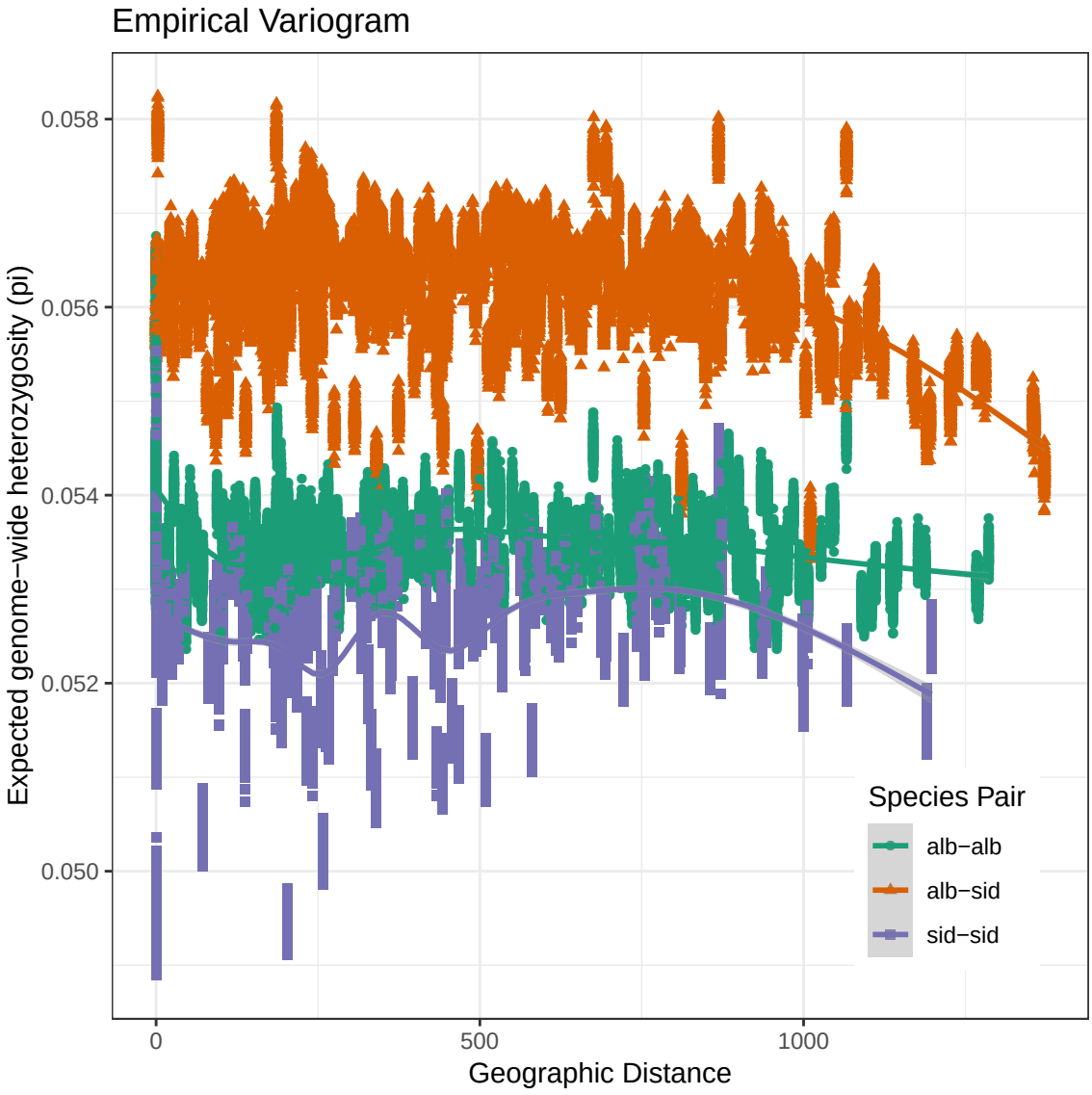



**Figure 6.14: Empirical variogram of intra- and inter-species locality comparisons.**

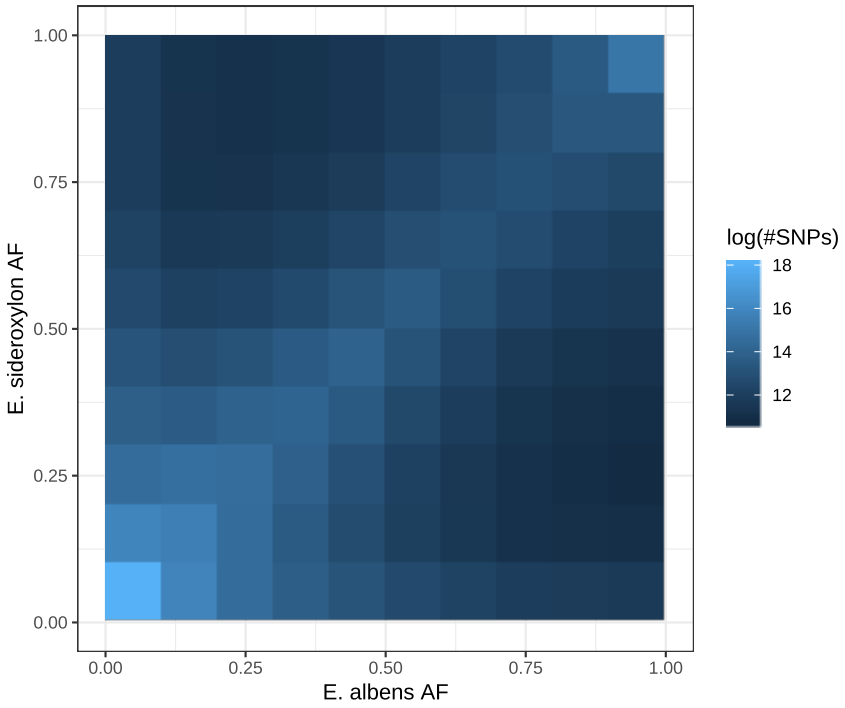



**Figure 6.15: Two-dimensional site frequency spectrum between *E. albens* and *E. sideroxylon*.**

**Table 6.3: GDM model variables, model deviance explained, and variable-specific p-values.**

| | | |
|---|---|---|
| *E. albens* | | |
| Variable | % dev. expl. | Variable p-value |
| Geographic | 31.475641 | 0 |
| Hydrologic conductivity - average saturated | 31.157494 | 0.05 |
| Precipitation - wettest month | 30.314927 | 0.02 |
| Precipitation - spring or autumn (log) season | 28.387513 | 0.13 |
| Phosphorus pre-European | 25.490866 | 0.19 |
| Wind run - month min | 22.645057 | 0.36 |
| *E. sideroxylon* | | |
| Variable | % dev. expl. | Variable p-value |
| Geographic | 89.887027 | 0 |
| Temperature - max difference in max | 89.599473 | 0 |
| Precipitation - max difference between successive months | 89.265338 | 0 |
| Vapour pressure deficit - month min | 87.875363 | 0 |
| Radiation - max month precipitation modified | 85.197611 | 0.04 |
| Water holding capacity - plant-available | 79.720207 | 0.12 |

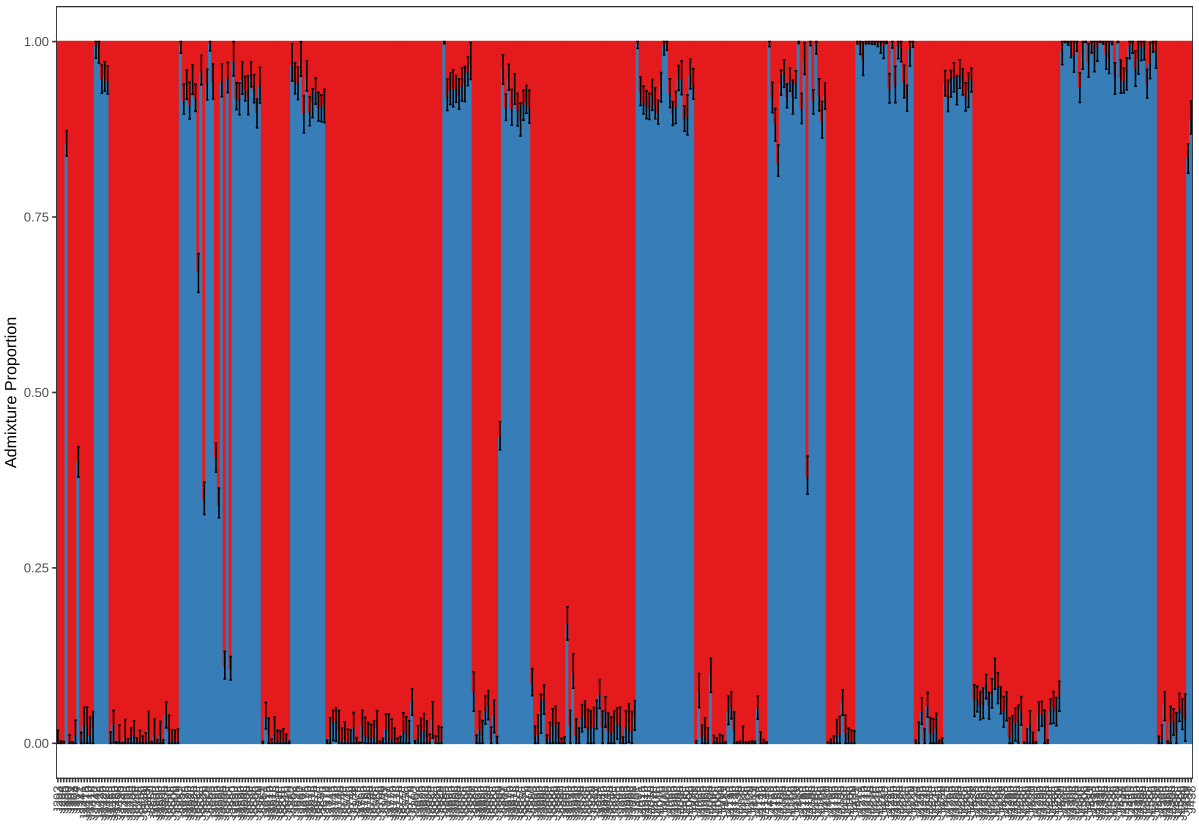



**Figure 6.16: Individual-level conStruct analysis of all samples with two model layers.** Admixture proportions are presented as means ± sd across 20 random subsets of 1 million SNPs. Note samples that appear as intermediates, suggesting they are recent interspecific hybrids.

# Chapter 7

# Thesis Discussion

## 7.1 Thesis progress

Through the work presented in this thesis, I have contributed both tools and knowledge to a greater understanding of the genomic variation across landscapes and the globe. I have developed new computational methods that enable researchers across the world to analyse genetic variation in datasets produced using modern genetic sequencing technology easily, accurately, and efficiently. Through my contribution to a global survey of genetic diversity in *Brachypodium*, we discover surprisingly low genetic variability and high inbreeding in these model species, and suggest further steps required to realise *Brachypodium* as a model for quantitative genetics. Through my examination of the spatial patterns of genetic diversity in woodland eucalypts, I describe the huge genetic diversity of those species, and find isolation by distance and environment landscape. These studies highlight the power of genomics to harness natural genetic variation to uncover the genetic basis of traits, and guide restoration of degraded ecosystems.

## 7.2 New computational methods

A large component of this thesis has been the development of new computational methods required for the analysis of genetic sequences. These methods range from low-level utilities to new methods implemented in large software projects and published in upper-tier bioinformatics journals. Of these tools, several are improved analysis methods for very high throughput Genotyping-by-sequencing (GBS) experiments. Axe (Chapter 2) enables the demultiplexing of many hundreds of samples from one sequencing run. Given the increased output of sequencing runs, multiplexing large numbers of samples is required in order to realise the

increased cost effectiveness of newer sequencers, and no demultiplexer was able to efficiently demultiplex index sequences required by the GBS protocol. The efficient and reusable implementations of low-level sequence quality control measures in libqcpp and trimit (Chapter 3) provide a "one-stop-shop" for both GBS and other data types, reducing the intellectual burden required to use a litany of tools for several similar steps of an analysis (particularly for those with less computational experience).

In developing kWIP (Chapter 4), I enable the estimation of genetic distance from raw sequencing data. This tool allows researchers to perform initial analyses like confirming sample identity or family structure without a reference genome. KWIP direct estimation of distance from sequence read data enables researchers to use downstream analyses that require only a genetic distance matrix (e.g., Generalised Dissimilarity Modelling of isolation by distance and environment) in situations where alignment to a reference genome is infeasible. However, genetic distance is only one of several metrics of interest in most studies. Using the same underlying data structures and approach employed in kWIP, one can estimate inter-sample covariance; in fact an approximation of sample covariance is an intermediate in kWIP. Many recent methods in population, landscape, and quantitative genomics define various processes as a function of sample covariance (e.g. population structure and IBD in conStruct Bradburd et al., 2017, geogenetic space in SpaceMix 2016; genomic selection using gBLUP VanRaden, 2008). Such a method would be especially powerful if it incorporated a PCAngsd-like optimisation of the covariance matrix (Meisner and Albrechtsen, 2018) to minimise the effect of sequencing noise with lower coverage datasets. Incorporating the estimation of sample covariance in a kWIP-like method would also assist some quantitative genomic methods, for example to extend gBLUP-based genomic selection (VanRaden, 2008). These genetic distance type association experiments can be extended to include microbiome information without the difficulty of metagenome assembly and are promising for prediction of traits in field samples.

The three methods I discuss in detail in this thesis are far from the only methods I have developed for analyses performed in this thesis. Currently unpublished tools include short-read sequencing utilities (e.g. `seqhax`; https://github.com/kdmurray91/seqhax), tools for retrieving data from the NCBI Sequence Read Archive (`srapy`; https://github.com/kdmurray91/srapy), and C++ and python libraries for exact and probabilistic kmer counting (`pymer` and `kmerkmer`; https://github.com/kmerkmer/). While in many cases these are far from the only tools to implement such functionality, they all advance previous implementations in some way (e.g. efficiency, ease of use, application specificity). These open source software

tools continue to be developed, and have been used in numerous projects, both within and beyond my immediate research environment. Additionally, I've contributed large quantities of code to external projects, which in some cases resulted in co-authorship on software publications (see Appendix A).

The aim of scientists developing new methods and software has predominantly been the creation of tools that provide accurate results, as this is overwhelmingly the concern of users (and rightly so). However, given the limited funding and scarcity of academic status given to those developing new software, these new methods often have less than ideal software quality and user friendliness, especially for methods developed outside one of only a few very well resourced bioinformatic method development teams (e.g. Broad Institute, Wellcome Trust, Sanger Centre). While such tools are obviously preferable to well-designed, user friendly, but methodologically inferior software, we need to acknowledge that the state of academic research software often imposes a significant intellectual burden on users. As an example, one software crucial to analyses performed in Chapter 6 had a bug which caused incorrect results to be produced from our dataset. To find and fix this bug, I required the ability to debug running C++ programs, and then deep knowledge of both C++ and the metric of interest. Without my background in software engineering this fix would not have been possible, and I would either have detected this error and terminated those analyses, or worse carried these incorrect outputs on to further analyses. I see it as the responsibility of tool developers to ensure their users need not be experts in software engineering to run their software successfully, a low bar that many bioinformatic methods fail to meet. Only through increased respect and funding for method development can we improve this situation.

In both Chapters 4 and 5, I demonstrate that the analysis of genomic data from leading-edge molecular methods can require the development of new software, often initially specific to some larger genomic project. This implies that direct collaboration between field biologists and computational biologists is just as important as the collaborations that typically exist between field biologists and the molecular biologists whose expertise generates these datasets. Going further, the training of field biologists in bioinformatics is no less important than their training in molecular biology. We must prevent the biologists of the future being either insufficiently confident or skilled computationally to analyse the large datasets that they will be able to generate with minimal expense. Good experimental design depends on knowledge across field, molecular, and computational biology, and when these skills are jointly brought to bear, exciting and now published results, methods, and findings have appeared.

## 7.3 Brachypodium as a model cereal

We performed a global survey of genetic diversity in the *Brachypodium distachyon* species complex, amassing a collection of over 3000 accessions (Chapter 5). Initial sequencing of this collection identified samples as different species, detected significant population structure within each species, and identified high levels of clonal family structure. We reduced the more than 800 accessions of *B. distachyon* to a core set of 74 accessions, and proceeded with a genome-wide association study (GWAS) for early vigour and energy use efficiency traits under both current and predicted future climate regimes. This association study found several significant QTL and high heritability. However, this GWAS had limited power due to the unexpectedly low genetic diversity of our global collection and resulting small core set of diverse accessions.

Previous authors reporting on smaller collections of Brachypodium from the native Iberian or middle eastern ranges reported relatively high genetic diversity. We expected our global collection to have similar genetic diversity to comparable model systems, e.g. *Arabidopsis* (Alonso-Blanco et al., 2016) and *Oryza sativa* (Li et al., 2014). Studies similar to ours in *Arabidopsis* have established GWAS sets that enabled the genetic basis of a wide variety of traits to be examined (e.g. Li et al., 2010, 2006; Alonso-Blanco et al., 2016; Atwell et al., 2010). While previous studies of *B. distachyon* did find population and family structure (e.g. Vogel et al., 2009), it was not to the extent observed here, perhaps as these studies focussed on the ancestral refugia that remain the hotspots of genetic diversity. The amount of genetic diversity identified here could be sufficient for GWAS, however, genetic lineages were not recombining and have nearly identical genomes. The only exceptions were within the ancestral hotspots of diversity in Turkey and Iberia (Brachi et al., 2011; Vogel et al., 2009). Therefore, further work crossing populations is required to perform powerful association studies in *Brachypodium*, specifically, the creation of mapping populations that disentangle genetic variation from background population structure across *B. distachyon* as a whole (for example creation of multiparental advanced-generation intercross, MAGIC, populations). Biparental mapping populations have been created previously (e.g. Vogel et al., 2009; Garvin et al., 2009; Huo et al., 2011).

Even without the creation of these mapping populations, *Brachypodium* remains an appealing model cereal for many areas of plant science (recently reviewed in Scholthof et al., 2018). The variation in ploidy and karotype within the *Brachypodium* species complex could be used to study the genomic effects of polyploidisation (Catalán et al., 2012). We identified individual clonal families present in over 40 localities including in both Australia and Turkey,

with significant variation in environment. Our findings have enabled an ongoing study of the possible epigenomic basis of this plasticity [Eichten et al. (2016); and ongoing studies]. Its small size and rapid life cycle make phenotyping *Brachypodium* in high throughput far easier than most crop species (e.g. wheat; Brkljacic et al., 2011). For this reason, *Brachypodium* is an attractive model for phenotyping-intensive studies such as forward-genetic mutation screens. We have made seed from our collection of over 2000 accessions available, and we have publicly deposited all associated data, enabling its use by the *Brachypodium* research community.

Model systems have been used to study an endless variety of questions across a large spectrum of biology. Traditionally, a major impediment to genomics in non-model species was the expense of establishing the genomic resources required for these studies. While still an expensive undertaking, this thesis demonstrates that technological advancement has made the resequencing of a moderately large set of diverse lines possible on a relatively modest budget, and the cost of long-read data for reference genome assembly has reduced markedly (e.g. Michael et al., 2018; Jain et al., 2018; Wu et al., 2019). This raises the question: are model systems even required? Having now worked across both model and non-model species, the appeal of model species remains. This economic feasibility now means that computational analysis is a bottleneck, and the resources available in model systems simplifies these analyses. The generation of whole-genome resequencing remains too expensive to be used for every genomic study in non-model species, and there is still a place for reduced-representation sequencing data. However, we are experiencing a dramatic improvement in the availability and quality of genomic resources for non-model and emerging model species, as can be seen from my work in *Eucalyptus* (Chapter 6).

## 7.4 Landscape drivers of eucalypt genetic diversity

In Chapter 6 I present a study of the patterns of genetic isolation across the range of two woodland eucalypt species. I found very high genetic diversity within each species, and relatively low differentiation between these species, with evidence of pervasive gene flow between them. We found no strong support for discrete population structure decoupled from Isolation By Distance (IBD). Whole-genome differentiation between localities was very low, though genomic distance correlates strongly to geographic distance in both species, particularly *E. sideroxylon*. There is an additional, though weaker, signal of IBE in each species, particularly driven by environmental variables describing the availability and timing of moisture, solar radiation, and soil nutrition.

### Gene flow, divergence, and species boundaries

Our findings highlight that a), eucalypt species exchange genetic material readily, and b) eucalypt species show low genome-wide divergence despite morphological differentiation. Evidence of the extent of reproductive isolation in *Eucalyptus* shows that inter-section, inter-series, and inter-specific gene flow occurs at nontrivial rates that are correlated with phylogenetic distance (Ellis et al., 1991; Larcombe et al., 2015). Speciation does not imply immediate whole-genome differentiation; it is possible that speciation is led by relatively few loci, and that speciation can occur in spite of ongoing gene flow (Nosil, 2008; Payseur and Rieseberg, 2016; Wu, 2001; e.g. in sunflower Andrew and Rieseberg, 2013; Ostevik et al., 2016). As outlined below, various experiments could probe the genetic underpinnings of reproductive isolation and speciation in these species.

Flowering time is an important pre-mating barrier to gene flow in *Eucalyptus* (Field et al., 2011). Reports of the flowering period of *E. albens* and *E. sideroxylon* differ, although there is a partial overlap in flowering time in most reports (Brooker and Kleinig, 2006; Costermans, 1983; Porter, 1978). Long-term flowering synchrony varied from very low to high across species pairs in a 30-year study of closely related box-ironbark species (Keatley et al., 2004), and synchrony varied year-to-year. The temporal barrier to gene flow between *E. albens* and *E. sideroxylon* is therefore unclear, but likely to be weak at most, though variable between years and across their common range. Validation of this hypothesis would have traditionally required prohibitively large-scale monitoring of flowering time across the range of both species, conducted over multiple years (e.g. Keatley et al., 2004). However, remote sensing imagery has proved an effective way of monitoring phenology at a massive scale (Zhang et al., 2003), and it is possible that it could be used to detect the extent of flowering in stands of these species across their ranges in the future (Nagendra et al., 2013; Viña et al., 2004).

A genome scan for loci strongly differentiated between *E. albens* and *E. sideroxylon* could identify loci driving the speciation process. Scanning the genome for loci that show much stronger differentiation between these species than the genome average could highlight loci of putative importance during speciation, although loci functionally irrelevant to speciation are also likely to be falsely identified (Strasburg et al., 2012). Natural hybrid zones present an ideal opportunity to discover the genetic basis of traits differentiated between species, as traits segregate independently in later generation hybrids (Gompert et al., 2017). Later generation hybrid seed from one or more *E. albens X sideroxylon* individuals could be treated as a multiparental mapping population, and used to dissect the genetic basis of traits fixed between species (see Pryor, 1953). Of particular interest are soil preference traits, with *E. albens*

known to prefer more fertile soils than *E. sideroxylon* (Brooker and Kleinig, 2006; Costermans, 1983). The genetic basis of seedling germination and establishment on different soils could be investigated in a hybrid zone. Such an experiment would address a potential cause of the observation that many closely-related *Eucalyptus* species have overlapping ranges yet are not commonly found at the same locality. Additionally, hybrid zones of these species could be used to investigate the genetic nature of the incomplete reproductive isolation between these species (Strasburg et al., 2012; e.g. in tomato Moyle and Nakazato, 2010; *Eucalyptus* Myburg et al., 2004).

Genome-wide differentiation between *E. albens* and *E. sideroxylon* is similar to levels of inter-population differentiation in some model species (e.g. *Arabidopsis*; Alonso-Blanco et al., 2016). Therefore, the use of methods that expect a shared pool of genetic variation and low inter-population differentiation (e.g. GWAS for traits segregating in both species) across closely related species in *Eucalyptus* should be feasible, although the use of sophisticated mixed-model based corrections for kinship would be particularly important in any such experiment given the expected strong structuring.

### Landscape drivers of genetic isolation

One surprising element of our results is the striking difference in the strength of IBD between *E. albens* and *E. sideroxylon*. Spatially autocorrelated intraspecific variation in flowering time would contribute to IBD, and differences in the strength of this spatial autocorrelation could be the basis of the differing strength of IBD in *E. albens* and *E. sideroxylon*. The stronger IBD observed in *E. sideroxylon* could also result of a supposed more historically discontinuous range (Costermans, 1983) and our data lend support to this theory.

We find support for modest isolation by environment (IBE) in both *E. albens* and *E. sideroxylon*, driven by availability and timing of moisture, solar radiation and soil nutrition. Broad-scale environmental factors such as regional climate may only explain a portion of the environmental isolation of such species. Environmental variation at finer spatial resolution likely also contributes to genetic isolation, for example, fine-scale terrain features that govern the availability of water (especially given the importance of broad-scale moisture availability we identified). Our sampling is aggregated to the locality-level, and these environmental variables are highly variable within localities; therefore, we cannot include these variables in our models of IBE despite their potential explanatory power. To test multi-scale hypotheses, one must adopt a more targeted sampling pattern, sampling over environmental gradients in a way that avoids confounding of environmental clines and broader isolating processes

like isolation by distance (Bragg et al., 2015; Wang and Bradburd, 2014). However, the genomic signals of fine-scale local adaptation may be inconsistent between localities, involving different causal variants (Ralph and Coop, 2010). Additionally, if selection is weak or variable through time, advantageous alleles are unlikely to sweep to fixation, instead allowing a diverse local gene pool to persist. High rates of pollen-mediated gene flow may counteract local adaptation, leading selection to filter out unsuitable genotypes in each generation. Where possible, experiments that investigate the genetic basis of this fine-scale adaptation are perhaps best performed in controlled experimental settings using early-life proxies of eventual fitness-related traits and controlled conditions that mimic environmental clines of interest.

Confounding among spatially and temporally autocorrelated environmental variables reduces statistical power, and risks increasing the false positive rate of any study that aims to correlate environment with genetics, either genome-wide or at specific loci. The genome-wide isolation by distance and environment observed in both *E. albens* and *E. sideroxylon* suggests attention must be paid to the issue of environmental autocorrelation in any genome-environment association studies on these species, even if inter-locality differentiation was low. Such confounding will always exist and, therefore, functional studies of phenotypic differentiation and local adaptation are required to confirm any association found in correlative experiments. However, such functional studies (e.g. replicated provenance trials) would be time-consuming and expensive at best in slowly maturing tree species, although functional studies on seedlings could provide a feasible solution for traits expressed early in development (e.g. Supple et al., 2018).

### Conservation implications

Landscape genomic experiments are increasingly used to guide conservation decisions, for example, guiding the selection of seed provenances that are predicted to be suited to future environments at a restoration locality (Broadhurst et al., 2008; Prober et al., 2015; Williams et al., 2014; e.g. in *E. melliodora* Supple et al., 2018). However, there are numerous pitfalls regarding the use of such results for these purposes (Kardos and Shafer, 2018). Firstly, genome-wide isolation by environment is not proof of local adaptation (Wang and Bradburd, 2014). Secondly, restoration using only "pre-adapted" seed risks limiting additional genetic diversity and therefore adaptive potential of restored populations (Broadhurst et al., 2008; Hoffmann et al., 2015), or introducing seed malapted in environmental axes not captured by landscape genomic models (e.g. migration across soil types could be suggested by models considering only climate). That restoration strategy also assumes that, a), models of IBE are accurate,

and b), the climate experienced by restored populations will be similar to the historic climate that influenced genetic isolation observed in these studies, and that models of future climate accurately predict weather that will be experienced by target populations in the future. The traditional practice of local provenancing brings its own genetic risks, particularly in highly fragmented species that exhibit strong inbreeding depression. Therefore, current best practice recommends a balance of local and climate-adjusted seed sources (Prober et al., 2015).

The current expense of population-wide genome resequencing studies for quantifying IBE, as in Chapter 6, makes their application to all species in need of conservation management unfeasible (Kardos and Shafer, 2018); emerging crop and foundation species must be prioritised. However, the genomic lessons learned in one species are likely to be sufficiently applicable to closely related species that restoration can be conducted with scarce genomic data, although more comparative studies are needed to understand how well models of IBD and IBE generalise across closely related species. Instead of starting with estimation of population structure, IBD, and/or IBE or a genotype-environment association study, one could directly assess the genetic filtering that occurs during establishment. To do so, we could revegetate numerous localities using a common, widely-collected, and diverse seed collection, using direct seeding or mass planting to minimise cost as per best practices (Broadhurst et al., 2008; Prober et al., 2015). Then, we could assess the genetic variants present among planted individuals that survive and thrive in each locality as time progresses. Genomics could be used as a diagnostic tool for distinguishing source localities or combinations of variants that appear maladapted. As well as their direct conservation benefits as high-quality restoration plantings, such experiments have the longer-term advantage of directly addressing important questions: which genetic backgrounds show highest fitness in which localities, and which underlying subsets of loci are associated with this increased survival in a given environment.

## 7.5 Evolutionary genomics in the Anthropocene

The Anthropocene, our current geological epoch that began around 1950, is characterised by human dominance over climate and the global environment (Steffen et al., 2011). Earth's ecosystems face an accelerating barrage of threats from humanity. Within Australia, landscape changes began 40,000-60,000 years before present with the first human settlement of the Australian continent. These impacts have included mass extinctions of animals (particularly large vertebrates), and drastic changes to fire regimes and vegetation types (Flannery, 1990; Miller et al., 1999; Roberts et al., 2001). Since European colonisation, human impacts

have been even more extreme, with widespread conversion of native habitats to urban and agricultural use (Bradshaw, 2012). Looking forward, ecosystems in Australia and elsewhere will continue to face numerous threats from human activity, including climate change, deforestation,land use conversion, and invasive species. We have begun to feel the unmistakable effects of anthropogenic changes to the global climate (Parmesan, 2006), and most predictions of future climates paint a grim picture for Australia's ecosystems and agricultural lands with general increases to temperature, increased weather variability, greater intensity and severity of extreme climate events (e.g., drought), loss of alpine habitats (Hughes, 2003), and negative effects on agriculture (Howden et al., 2007). While anthropogenic climate change poses a major threat to ecosystems, continued land clearing in many cases is a more immediate threat, and may obliterate some habitats before climate change has the chance.

Genomics has a role to play in solving these interconnected challenges but requires a unified view of Earth's natural and anthropogenically-modified ecosystems. In particular, we must advance technological, agronomic, and social solutions that increase crop yields while restoring degraded lands and ecosystems in the long term. Increased crop yields are attainable, and will be required to provide a healthy diet for a developing human population (Willett et al., 2019), as further expansion of agricultural land comes with high environmental costs (Garnett et al., 2013; Ray et al., 2013; Tilman et al., 2011). Current management of agricultural land is often woefully suboptimal in environmental and economic terms (Beyer et al., 2018; Foley et al., 2011). Yet, this leaves substantial room for improvement. Addressing agronomic shortcomings via technological advancements within existing agricultural systems can reduce further harms from continuing agricultural expansion. Reducing expansion is insufficient: we must also restore currently degraded ecosystems to enhance their ecological functions such that they are more resilient to extreme weather and economic shocks. So-called "regenerative" agricultural practices (e.g. agroforestry, perennial cropping) promise an economically and socially scalable solution that can provide both food and habitat. Consistently poor agricultural land could be re-forested to provide ecosystem services, while genome-assisted selection of crops could increase inter-cropping yield on remaining arable land (Rivers et al., 2015). These strategies show promise, although further development is needed for them to become economically scaleable (reviewed in Toensmeier, 2016). The threats of climate change, loss of biodiversity and ecosystem function, and global food shortages are hard to understate. Most solutions require at least as much social and behavioural change as technological change and scientific discovery. Advancing research into the synergistic benefits of agroforestry and genome-assisted crop selection is a necessary step towards real solutions.

Evolutionary and ecological genomics are well-poised to assist these challenging feats. Focusing on foundation species that provide critical ecosystem functions and key underdeveloped food crops is likely to be the most effective strategy. A focus on foundation species would provide umbrella benefits to a much larger community of associated species, thereby maximising limited time and funding resources. Efficiently measuring extant (phylo-)genetic diversity across a wide range of species would serve as a benchmark by which we can assess the effectiveness of any action. Accurate determination of past evolution to extreme climates is likely important, in crops, crop wild relatives, and foundation species. An evolutionary or ecological perspective on crop breeding has proved highly effective, for example successful introgression of pathogen resistance from crop wild relatives (Denison, 2017; Hajjar and Hodgkin, 2007; Piquerez et al., 2014). Genomic methods will accelerate domestication of new crop species and can select for traits that have large positive environmental impact to provide both ecosystem services and human food.

In summary, evolutionary genomics can assist a wide range of urgent and pending questions relating to our global land use. It can unleash resilient new feed, fiber, and fodder crops, and inform management of ecosystems to address health and environmental security for the 21st century.

## 7.6 References


Alonso-Blanco C et al. **2016**. "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana." *Cell* **166**:481–491. doi:10.1016/j.cell.2016.05.063

Andrew RL, Rieseberg LH. **2013**. "Divergence Is Focused on Few Genomic Regions Early in Speciation: Incipient Speciation of Sunflower Ecotypes." *Evolution* **67**:2468–2482. doi:10.1111/evo.12106

Atwell S et al. **2010**. "Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines." *Nature* **465**:627–631. doi:10.1038/nature08800

Beyer RM, Manica A, Rademacher TT. **2018**. "Relocating agriculture could drastically reduce humanity's ecological footprint." *bioRxiv* 488841. doi:10.1101/488841

Brachi B, Morris GP, Borevitz JO. **2011**. "Genome-wide association studies in plants: The missing heritability is in the field." *Genome Biol* **12**:232. doi:10.1186/gb-2011-12-10-232

Bradburd G, Coop G, Ralph P. **2017**. "Inferring Continuous and Discrete Population Genetic Structure Across Space." *bioRxiv* 189688. doi:10.1101/189688

Bradburd GS, Ralph PL, Coop GM. **2016**. "A Spatial Framework for Understanding Pop-

ulation Structure and Admixture." *PLOS Genetics* **12**:e1005703. doi:10.1371/journal.pgen.1005703

Bradshaw CJA. **2012**. "Little left to lose: Deforestation and forest degradation in Australia since European colonization." *Journal of Plant Ecology* **5**:109–120. doi:10.1093/jpe/rtr038

Bragg JG, Supple MA, Andrew RL, Borevitz JO. **2015**. "Genomic variation across landscapes: Insights and applications." *New Phytol* **207**:953–967. doi:10.1111/nph.13410

Brkljacic J et al. **2011**. "Brachypodium as a Model for the Grasses: Today and the Future." *Plant Physiology* **157**:3–13. doi:10.1104/pp.111.179531

Broadhurst LM, Lowe A, Coates DJ, Cunningham SA, McDonald M, Vesk PA, Yates C. **2008**. "Seed supply for broadscale restoration: Maximizing evolutionary potential." *Evolutionary Applications* **1**:587–597. doi:10.1111/j.1752-4571.2008.00045.x

Brooker I, Kleinig D. **2006**. "Field guide to eucalypts." Melbourne: Bloomings Books.

Catalán P et al. **2012**. "Evolution and taxonomic split of the model grass Brachypodium distachyon." *Ann Bot* **109**:385–405. doi:10.1093/aob/mcr294

Costermans L. **1983**. "Native trees and shrubs of south-eastern Australia." Sydney: Reed.

Denison RF. **2017**. "Darwinian agriculture: How understanding evolution can improve agriculture."

Eichten SR, Stuart T, Srivastava A, Lister R, Borevitz JO. **2016**. "DNA methylation profiles of diverse Brachypodium distachyon aligns with underlying genetic diversity." *Genome Res* gr.205468.116. doi:10.1101/gr.205468.116

Ellis MF, Sedgley M, Gardner JA. **1991**. "Interspecific Pollen-Pistil Interaction in Eucalyptus L'Hér. (Myrtaceae): The Effect of Taxonomic Distance." *Ann Bot* **68**:185–194. doi:10.1093/oxfordjournals.aob.a088243

Field DL, Ayre DJ, Whelan RJ, Young AG. **2011**. "The importance of pre-mating barriers and the local demographic context for contemporary mating patterns in hybrid zones of Eucalyptus aggregata and Eucalyptus rubida." *Molecular Ecology* **20**:2367–2379. doi:10.1111/j.1365-294X.2011.05054.x

Flannery TF. **1990**. "Pleistocene faunal loss: Implications of the aftershock for Australia's past and future." *Archaeology in Oceania* **25**:45–55. doi:10.1002/j.1834-4453.1990.tb00232.x

Foley JA et al. **2011**. "Solutions for a cultivated planet." *Nature* **478**:337–342. doi:10.1038/nature10452

Garnett T et al. **2013**. "Sustainable Intensification in Agriculture: Premises and Policies." *Science* **341**:33–34. doi:10.1126/science.1234485

Garvin DF et al. **2009**. "An SSR-based genetic linkage map of the model grass Brachypodium distachyon." *Genome* **53**:1–13. doi:10.1139/G09-079

Gompert Z, Mandeville EG, Buerkle CA. **2017**. "Analysis of Population Genomic Data from Hybrid Zones." *Annu Rev Ecol Evol Syst* **48**:207–229. doi:10.1146/annurev-ecolsys-110316-022652

Hajjar R, Hodgkin T. **2007**. "The use of wild relatives in crop improvement: A survey of developments over the last 20 years." *Euphytica* **156**:1–13. doi:10.1007/s10681-007-9363-0

Hoffmann A et al. **2015**. "A framework for incorporating evolutionary genomics into biodiversity conservation and management." *Climate Change Responses* **2**:1. doi:10.1186/s40665-014-0009-x

Howden SM, Soussana J-F, Tubiello FN, Chhetri N, Dunlop M, Meinke H. **2007**. "Adapting agriculture to climate change." *PNAS* **104**:19691–19696. doi:10.1073/pnas.0701890104

Hughes L. **2003**. "Climate change and Australia: Trends, projections and impacts." *Austral Ecology* **28**:423–443. doi:10.1046/j.1442-9993.2003.01300.x

Huo N, Garvin DF, You FM, McMahon S, Luo M-C, Gu YQ, Lazo GR, Vogel JP. **2011**. "Comparison of a high-density genetic linkage map to genome features in the model grass Brachypodium distachyon." *Theor Appl Genet* **123**:455–464. doi:10.1007/s00122-011-1598-4

Jain M et al. **2018**. "Nanopore sequencing and assembly of a human genome with ultra-long reads." *Nature Biotechnology* **36**:338–345. doi:10.1038/nbt.4060

Kardos M, Shafer ABA. **2018**. "The Peril of Gene-Targeted Conservation." *Trends in Ecology & Evolution* **33**:827–839. doi:10.1016/j.tree.2018.08.011

Keatley MR, Hudson IL, Fletcher TD. **2004**. "Long-term flowering synchrony of box-ironbark eucalypts." *Aust J Bot* **52**:47–54. doi:10.1071/bt03017

Larcombe MJ, Holland B, Steane DA, Jones RC, Nicolle D, Vaillancourt RE, Potts BM. **2015**. "Patterns of Reproductive Isolation in Eucalyptus—A Phylogenetic Perspective." *Mol Biol Evol* **32**:1833–1846. doi:10.1093/molbev/msv063

Li J-Y, Wang J, Zeigler RS. **2014**. "The 3,000 rice genomes project: New opportunities and challenges for future rice research." *GigaScience* **3**:8. doi:10.1186/2047-217X-3-8

Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO. **2010**. "Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana." *PNAS* **107**:21199–21204. doi:10.1073/pnas.1007431107

Li Y, Roycewicz P, Smith E, Borevitz JO. **2006**. "Genetics of Local Adaptation in the Laboratory: Flowering Time Quantitative Trait Loci under Geographic and Seasonal Conditions in Arabidopsis." *PLoS ONE* **1**:e105. doi:10.1371/journal.pone.0000105

Meisner J, Albrechtsen A. **2018**. "Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data." *Genetics* **210**:719–731. doi:10.1534/genetics.118.301336

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. **2018**. "High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell." *Nature Communications* **9**:541. doi:10.1038/s41467-018-03016-2

Miller GH, Magee JW, Johnson BJ, Fogel ML, Spooner NA, McCulloch MT, Ayliffe LK. **1999**. "Pleistocene Extinction of Genyornis newtoni: Human Impact on Australian Megafauna." *Science* **283**:205–208. doi:10.1126/science.283.5399.205

Moyle LC, Nakazato T. **2010**. "Hybrid Incompatibility 'Snowballs' Between Solanum Species." *Science* **329**:1521–1523. doi:10.1126/science.1193063

Myburg AA, Vogl C, Griffin AR, Sederoff RR, Whetten RW. **2004**. "Genetics of Postzygotic Isolation in Eucalyptus: Whole-Genome Analysis of Barriers to Introgression in a Wide Interspecific Cross of Eucalyptus grandis and E. Globulus." *Genetics* **166**:1405–1418. doi:10.1534/genetics.166.3.1405

Nagendra H, Lucas R, Honrado JP, Jongman RHG, Tarantino C, Adamo M, Mairota P. **2013**. "Remote sensing for conservation monitoring: Assessing protected areas, habitat extent, habitat condition, species diversity, and threats." *Ecological Indicators*, Biodiversity Monitoring **33**:45–59. doi:10.1016/j.ecolind.2012.09.014

Nosil P. **2008**. "Speciation with gene flow could be common." *Molecular Ecology* **17**:2103–2106. doi:10.1111/j.1365-294X.2008.03715.x

Ostevik KL, Andrew RL, Otto SP, Rieseberg LH. **2016**. "Multiple reproductive barriers separate recently diverged sunflower ecotypes." *Evolution* **70**:2322–2335. doi:10.1111/evo.13027

Parmesan C. **2006**. "Ecological and Evolutionary Responses to Recent Climate Change." *Annual Review of Ecology, Evolution, and Systematics* **37**:637–669. doi:10.1146/annurev.ecolsys.37.091305.110100

Payseur BA, Rieseberg LH. **2016**. "A genomic perspective on hybridization and speciation." *Molecular Ecology* **25**:2337–2360. doi:10.1111/mec.13557

Piquerez SJM, Harvey SE, Beynon JL, Ntoukakis V. **2014**. "Improving crop disease resistance: Lessons from research on Arabidopsis and tomato." *Front Plant Sci* **5**. doi:10.3389/fpls.2014.00671

Porter JW. **1978**. "Relationships between flowering and honey production of red ironbark, Eucalyptus sideroxylon (A. Cunn.) Benth., And climate in the Bendigo district of Victoria." *Aust J Agric Res* **29**:815–829. doi:10.1071/ar9780815

Prober SM, Byrne M, McLean EH, Steane DA, Potts BM, Vaillancourt RE, Stock WD. **2015**. "Climate-adjusted provenancing: A strategy for climate-resilient ecological restora-

tion." *Front Ecol Evol* **3**. doi:10.3389/fevo.2015.00065

Pryor LD. **1953**. "Anther shape in Eucalyptus genetics and systematics." *Proceedings of the Linnean Society of New South Wales* **78**:43–48.

Ralph P, Coop G. **2010**. "Parallel Adaptation: One or Many Waves of Advance of an Advantageous Allele?" *Genetics* **186**:647–668. doi:10.1534/genetics.110.119594

Ray DK, Mueller ND, West PC, Foley JA. **2013**. "Yield Trends Are Insufficient to Double Global Crop Production by 2050." *PLOS ONE* **8**:e66428. doi:10.1371/journal.pone.0066428

Rivers J, Warthmann N, Pogson BJ, Borevitz JO. **2015**. "Genomic breeding for food, environment and livelihoods." *Food Sec* **7**:375–382. doi:10.1007/s12571-015-0431-3

Roberts RG et al. **2001**. "New Ages for the Last Australian Megafauna: Continent-Wide Extinction About 46,000 Years Ago." *Science* **292**:1888–1892. doi:10.1126/science.1060264

Scholthof K-BG, Irigoyen S, Catalan P, Mandadi K. **2018**. "Brachypodium: A monocot grass model system for plant biology." *The Plant Cell* tpc.00083.2018. doi:10.1105/tpc.18.00083

Steffen W, Grinevald J, Crutzen P, McNeill J. **2011**. "The Anthropocene: Conceptual and historical perspectives." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **369**:842–867. doi:10.1098/rsta.2010.0327

Strasburg JL, Sherman NA, Wright KM, Moyle LC, Willis JH, Rieseberg LH. **2012**. "What can patterns of differentiation across plant genomes tell us about adaptation and speciation?" *Philos Trans R Soc Lond B Biol Sci* **367**:364–373. doi:10.1098/rstb.2011.0199

Supple MA, Bragg JG, Broadhurst LM, Nicotra AB, Byrne M, Andrew RL, Widdup A, Aitken NC, Borevitz JO. **2018**. "Landscape genomic prediction for restoration of a Eucalyptus foundation species under climate change." *eLife* **7**:e31835. doi:10.7554/eLife.31835

Tilman D, Balzer C, Hill J, Befort BL. **2011**. "Global food demand and the sustainable intensification of agriculture." *PNAS* **108**:20260–20264. doi:10.1073/pnas.1116437108

Toensmeier E. **2016**. "The Carbon Farming Solution: A Global Toolkit of Perennial Crops and Regenerative Agriculture Practices for Climate Change Mitigation and Food Security." Chelsea Green Publishing.

VanRaden PM. **2008**. "Efficient Methods to Compute Genomic Predictions." *Journal of Dairy Science* **91**:4414–4423. doi:10.3168/jds.2007-0980

Viña A, Gitelson AA, Rundquist DC, Keydan G, Leavitt B, Schepers J. **2004**. "Monitoring Maize ( Zea mays L.) Phenology with Remote Sensing." *Agronomy Journal* **96**:1139–1147. doi:10.2134/agronj2004.1139

Vogel JP, Tuna M, Budak H, Huo N, Gu YQ, Steinwand MA. **2009**. "Development of

SSR markers and analysis of diversity in Turkish populations of Brachypodium distachyon." *BMC Plant Biology* **9**:88. doi:10.1186/1471-2229-9-88

Wang IJ, Bradburd GS. **2014**. "Isolation by environment." *Mol Ecol* **23**:5649–5662. doi:10.1111/mec.12938

Willett W et al. **2019**. "Food in the Anthropocene: The EAT–Lancet Commission on healthy diets from sustainable food systems." *The Lancet* **0**. doi:10.1016/S0140-6736(18)31788-4

Williams AV, Nevill PG, Krauss SL. **2014**. "Next generation restoration genetics: Applications and opportunities." *Trends in Plant Science* **19**:529–537. doi:10.1016/j.tplants.2014.03.011

Wu C-I. **2001**. "The genic view of the process of speciation." *Journal of Evolutionary Biology* **14**:851–865. doi:10.1046/j.1420-9101.2001.00335.x

Wu M, Kostyun JL, Moyle LC. **2019**. "Genome sequence of Jaltomata addresses rapid reproductive trait evolution and enhances comparative genomics in the hyper-diverse Solanaceae." *Genome Biol Evol*. doi:10.1093/gbe/evy274

Zhang X, Friedl MA, Schaaf CB, Strahler AH, Hodges JCF, Gao F, Reed BC, Huete A. **2003**. "Monitoring vegetation phenology using MODIS." *Remote Sensing of Environment* **84**:471–475. doi:10.1016/S0034-4257(02)00135-9

# Appendices

# Appendix A

# Other published works

Aside from the works presented in chapters 2-6, I have worked on numerous projects that have lead to peer-reviewed publications during my PhD candidature. Below, I list and briefly describe these works.

- MR Crusoe *et al.* (2015): The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research 4
  - I contributed a large quantity of code to khmer, a software package that I used as a dependency of kWIP (Chapter 4).
- LC Teasdale *et al.* (2016): Identification and qualification of 500 nuclear, singlecopy, orthologous genes for the Eupulmonata (Gastropoda) using transcriptome sequencing and exon capture. Molecular ecology resources 16 (5), 1107-1123
  - I performed several phylogenetic analyses presented in this paper
- PA Crisp *et al.* (2017): Rapid recovery gene downregulation during excess-light stress and recovery in Arabidopsis The Plant Cell, 00828.2016 12
  - I developed two novel software tools for RNA degradome analysis, as we as performed most of statistical modelling of RNA decay in this paper.
- D Standage *et al.* (2017): khmer release v2. 1: software for biological sequence analysis. The Journal of Open Source Software 2 (15), 272
  - A continuation of my contribution of code to the khmer software package.
- PA Crisp *et al.* (2018): RNA Polymerase II read-through promotes expression of neighboring genes in SAL1-PAP-XRN retrograde signaling. Plant physiology 178 (4), 1614-1630

– I contributed to the design and execution of the statistical analysis of downstream gene expression presented in this paper.